STATISTICAL MODELS OF INCOMPLETE DATA AND THEIR USE IN THE SOCIAL SCIENCES

Ulrich Pötter

Habilitationsschrift Bochum 2008

Preface

The present document has been submitted as the author's *Habilitationsschrift* to the faculty of Social Science at the Ruhr University Bochum. It is concerned with the problem of incomplete data in the social sciences.

While traditionally the focus of statistical treatments of the problem was confined to completely missing data, the extensions to partially missing values were pursued systematically only since about 15 years. This approach has opened the way to unify concepts and models from otherwise quite divergent fields. Some problems of model building and estimation within the general approach are documented in the following pages with particular emphasis on social science applications.

The document is based on six articles that appeared either in refereed journals or as refereed contributions to conference proceedings. The six articles are:

- Probabilistische Selektionsmodelle, Kölner Zeitschrift f. Soziologie u. Sozialpsychologie, Sonderheft 44 "Methoden der Sozialforschung" (A. Diekmann, Ed.), 2006, 172–202
- Causal inference from series of events, European Sociological Review, 17, 2001, 21–32 (with H.-P. Blossfeld)
- A non-parametric mean residual life estimator, Metodoloski Zvezki, 19, 2003, 97–113 (with K. Kopperschmidt)
- Covariate effects in periodic hazard rate models, Metodoloski Zvezki, 17, 2002, 137–145 (with K. Kopperschmidt)
- A multivariate Buckley–James estimator, in: T. Kollo, E.-M. Tiit, M. Srivastava (Eds.): Multivariate Statistics, Utrecht 2000

- On the proportionality of regression coefficients in misspecified general linear models, in: E.-M. Tiit et al. (Eds.): Proceedings of the 5th Tartu Conference on Multivariate Statistics, Utrecht 1995

These articles are slightly revised to provide for a more unified notation. However, I have tried not to change the style or the arguments originally presented. All articles are followed by a new postscriptum highlighting later developments and putting them in the context of the general discussion of incomplete data.

The articles are preceded by two new chapters. The first chapter systematically introduces the theory of incomplete data. It surveys the probabilistic theory and highlights those features that are of critical importance for social science applications. The main advantages of a general formulation of the problem of incomplete data are presented and some alternative formulations are discussed.

A second new chapter presents a case study that serves both to illustrate the basic concepts and to emphasise those aspects of the theory that are of particular importance for social science applications. Also added is an appendix that treats some implementation problems of one of the central building blocks of the treatment of incomplete data, the estimation of distribution functions from incomplete data.

Contents

1

Introduction

Since 30 years or more, the literature on incomplete data is dominated by approaches that make essential use of statistical models and their underlying probabilistic desiderata: random variables, distributions, stochastic independence etc. In their famous book "Statistical Analysis with Missing Data", Roderick Little and Donald Rubin state:

> Missing data mechanisms are crucial since the properties of missing-data methods depend very strongly on the nature of the dependencies in these mechanisms. The crucial role of the mechanism in the analysis of data with missing values was largely ignored until the concept was formalized in the theory of Rubin (1976), through the simple device of treating the missing-data indicators as random variables and assigning them a distribution. (Little, Rubin 2002: 11)

This "simple device" is now nearly universally adopted both in theoretical research and in applied work, and particularly in empirical social research. This document is no exception and in effect most chapters that follow use the "simple device" without further justification.

However, this "simple device" is neither the most natural nor the only one available to the empirically working social scientist. Nor was the "simple device" accepted without resistance. In fact, it has often been doubted that a statistical treatment of the problem is at all possible, at least in the social sciences. Oskar Morgenstern, in his early study "On the Accuracy of Economic Observations", said that "the errors in these data cannot always be formulated according to strict statistical theory for the simple reason that no such exhaustive theory is available for many social phenomena" (1963: 7). Some 20 years later, Tore Dalenius commented on the statistical treatment of survey non-response:

> I take a dim view of the usefulness of these endeavors on two grounds. (1) First, it appears utterly unrealistic to postulate "response probabilities" which are independent of the varying circumstances under which an effort is made to elicit a response. ... (2) ... it seems unavoidable to introduce assumptions of unknown validity about probabilities. In summary, I am inclined to reject approaches to the non-response problem which involve "response probabilities" (Dalenius in Madow, Olkin 1983, vol. 3: 412).

While the resistance against statistical treatments of incomplete data has dwindled away, this may only partly be due to the successes of the the "simple device" of Little and Rubin. It may as well be due to the rather naive hope for a "method" that could somehow amend the deficiencies of data sets. The "simple device" plus a few "plausible assumptions" will indeed allow for estimates nearly as precise as the ones one would expect when problems of incomplete data were completely ignored.

In contrast, alternative approaches are much more cautious and will in many cases even indicate the inadequacy of a given data set to provide useful answers at all. Manski (2003: 18) illustrates the effect by looking at the percentage of employed persons estimated from the National Longitudinal Survey of Youth, a sample of 6812 young Americans. The naive estimate using only complete data is 78% with a 95% confidence interval of roughly ±1%. Both the estimate and the confidence interval can be justified within the Little-Rubin setup by simply declaring that the "missing data mechanism" is "missing at random" (Little, Rubin 2002: 12). But taking into account that people who did not answer to this question (some 18%) might be either employed or not, the percentage of employed may vary between 64% and 82%, an interval not only much larger than the one suggested by the sampling distribution but also too large for many practical purposes.

The computation is easy: For 82% of the interviewed the employment status is known, but for the other 18% it is not. Thus if all those not answering that question were unemployed or out of labour force, the proportion of employed in the sample would be $0.82 \cdot 0.78 + 0.18 \cdot 0 = 0.6396$ or roughly 64%. On the other hand, all of those not answering might as well have been employed. In that case, the share of 18% not answering must be added to the 64%, resulting in 82%. Note that these two extreme possibilities are not only logically consistent with the data but that they cannot be ruled out by an appeal to the general statistical approach to missing data nor by the more special notion of "missing at random". The only information available from the data alone is that the percentage of employed people in the sample is in the range of 64% to 82%. In consequence, any "inference" to the population at large must at least contain this interval, plus any uncertainty resulting from the sampling procedure.

It is obvious why most practitioners would prefer to report 78% plus minus 1% as the result of the survey instead of being forced to admit that the sought for percentage is somewhere between 64% and 82%. Conventional sampling theory suggests that in order to achieve a 95% (symmetric) confidence interval for the percentage of $\pm 9\%$ (the equivalent to the interval of 64% to 82% resulting from Manski's bounds), a sample of size 125 would be sufficient. Why go through the trouble to interview 6812 people if the result was at most as precise as the answers of just 125 persons? Moreover, since of the 6812 interviewed persons 5556 actually gave an answer, should the survey not be counted as being as informative as a pure random sample based on 5556 persons without any missing values?

In summary, there are many reasons to prefer the statistical approach to incomplete data, especially in the social sciences, where non-response rates of 18% must be regarded as moderate. But neither convenience, nor cost, nor appeal to conventional wisdom in sampling theory can be a valid argument to the effect that 78% $\pm 1\%$ is a reasonable estimate of the employment rate. How can such an optimistically accurate estimation be justified? The statistical approach to incomplete data proceeds in several steps:

The most crucial step is to construct a special mathematical model to replace the simple question originally posed. To make this step more transparent and to prepare for some generalisations, some notation is needed: Let $\mathcal{U} = \{u_1, \dots, u_N\}$ denote (a list of) the population of interest, in the example Americans aged 25 to 35 in 1991. Let

$$Y : \mathcal{U} \longrightarrow \{0, 1\} =: \mathcal{Y}$$

be a function that assigns the employment status (1 stays for 'employed', 0 for 'not employed') to all members of the population. Further, denote by $S \subseteq \mathcal{U}$ the given survey sample. The sought for percentage of employed people in the sample and in the population, respectively, are then

$$m(Y, S) := \frac{1}{|S|} \sum_{u \in S} Y(u) \qquad m(Y, \mathcal{U}) := \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} Y(u)$$

But some of the people did not answer to the question on their employment status and thus the range of Y must be enlarged to allow for this possibility. The function

$$Y^* : \mathcal{U} \longrightarrow \{\{0\}, \{1\}, *\} \cup \{\mathcal{Y}\} = \{\{0\}, \{1\}, \{0, 1\}, *\} =: \mathcal{Y}^*$$

will describe the answer to the employment question, where a '$*$' is used if people were not asked (were never in the survey) and a non-response is indicated by the set $\{0, 1\}$: a non-respondent might be either employed or not and the union of the two possibilities is the set $\{0, 1\}$.

The function Y^* is intended to capture the responses of sample members, while Y captures the (socially constructed) fact of employment. The latter is defined for essentially all members of the population. But an answer to a survey questions is not, unless one was willing to posit that all members of the population posses an "answer characteristic", assumed fixed for the purposes of survey sampling, that they "activate" just in case that they are included in a survey. The symbol '$*$' is introduced so that Y^* can still be defined for all members of the population without presupposing anything about "answer characteristics". This construction may be needed even when no comparison with some population is contemplate. E.g., the National Longitudinal Survey of Youth was started in 1979 as a panel study. Of the 1256 non-responses on the employment question in 1991, only 555 were approached in 1991 and either declined to be interviewed or did not answer to the question. The other 701 could not be located or were otherwise lost to follow up before 1991. To presuppose a certain “answer characteristic” and thus a particular value of $\{0\}$, $\{1\}$, or $\{0, 1\}$ for them would be preposterous and an extra symbol like ‘$*$’ should be introduced.

Allowing for non-response and thus using only the values of Y^* (restricted to the survey members), the value of $m(Y, S)$ must be between

$$\frac{1}{|S|}\left(\sum_{u \in S} \mathbb{1}[Y^* = \{1\}](u)\right) \text{ and } \frac{1}{|S|}\left(\sum_{u \in S} \mathbb{1}[Y^* \in \{\{1\}, \{0,1\}\}](u)\right) \tag{1.1}$$

where $\mathbb{1}[.]$ denotes the indicator function.[1] These are, of course, the bounds derived earlier. Note that there is nothing in the data that would allow to narrow the bounds. Both extreme cases are logically and materially possible, since the only connection between the functions is that from $Y^*(u) = \{1\}$ $Y(u) = 1$ should follow, and similarly for $Y^*(u) = \{0\}$.[2] The latter conditions stipulate that respondents tell the truth in that generally $Y(u) \in Y^*(u)$. This will be assumed throughout the following pages, since otherwise there would be no necessary connection between the observations and the percentage of employed in the sample. In the general situation, the requirement $Y(u) \in Y^*(u)$ will be termed *consistency condition* in the following.

The “simple device”, then, consists in adding some further structure to the setup. It does so by introducing a probability model: Let $(\Omega, \mathcal{B}, \lambda)$ be a probability space. As is usual in the construction of probability models, the nature of $(\Omega, \mathcal{B})$ is irrelevant if only the space Ω and its $\sigma-$algebra $\mathcal{B}$

[1] As usual, the indicator function $\mathbb{1}[A](x)$ takes the value 1 iff $x \in A$, 0 otherwise. For a function f, $\mathbb{1}[f \in B](x)$ is short for the somewhat cumbersome $\mathbb{1}[f^{-1}(B)](x)$ where $f^{-1}(B)$ is the pre-image of B.

[2] Note that the problem of “inference to the population” does not play any role in this setup. It will re-surface when sampling properties of the “simple device” are discussed.

is chosen large enough to allow for the following constructions. Now the functions Y and Y^* can be transformed into random variables by redefining them to have joint domain $\mathcal{U} \times \Omega$:[3]

$$(Y, Y^*)\colon \mathcal{U} \times \Omega \longrightarrow \mathcal{Y} \times \mathcal{Y}^*$$

These new functions should be carefully distinguished from the ones previously defined. The latter are a simple bookkeeping device, referring to a listing of people together with their employment status and their answers to survey questions respectively. In contrast, the new versions are only defined within a probability model. They are a mathematical construction and as such do not refer to anything in the world except in the very limited sense that it may be assumed that there is an ω such that the values of $Y(., \omega)$ are equal to the values in the list of the population. Values of both Y and Y^* are assumed to be generated from the list $\mathcal{U}$ together with draws from a random number generator that produces both values for each $u \in \mathcal{U}$. This is certainly not the way in which either the employment status or answers in interviews come about in this world.

Nevertheless it is exactly the formulation of a mathematical model that is the main achievement of the statistical approach. Within the model it is possible to precisely state conditions that justify a point estimate of a 78% employment rate. In terms of the probability model itself, the condition uniquely identifies the distribution of Y from the (known) distribution of Y^*. This condition is termed *missing at random* (MAR) or, in a more general context, *coarsening at random* (CAR). It states that the conditional probability of any set $\{Y^* = y^*\}$ given $\{Y = y\}$ for a particular value of y is the same for all $y \in y^*$. Put another way, the conditional probability of $\{Y = y\}$ given the set $\{Y^* = y^*\}$ is the same as that of the conditional probability of $\{Y = y\}$ given that $\{Y \in y^*\}$. In other words, having "observed" $\{Y^* = y^*\}$ this tells us that $\{Y \in y^*\}$ and nothing more. In Bayesian terms, the "information" $\{Y^* = y^*\}$ allows

[3] In Chapter 4 I have chosen to make the sample S dependent on $\omega \in \Omega$ as well so that one can treat questions of sampling design within the same model. However, for the purpose of this introduction such a generalisation may distract from the main argument. Thus, the sample S is taken to be a known and fixed set here.

to update the a priori distribution of Y to the conditional probability $\Pr(Y = y \mid Y \in y^*)$.

In the case of employment status the latter formulation states that the probability of being employed among the non-responders must be the same as the probability of being employed. Thus the probability of being employed, say p_1, must be equal to the probability of a positive answer to the employment question (0.64) plus the fraction of employed among the non-respondents ($0.18\, p_1$). This gives the probability of 78% employment.

This looks like a rather complicated reformulation of a simple solution that moreover could have been formulated without the help of the "simple device" and without any probability theory. Namely: the proportion of employed among the respondents is the same as that among the non-respondents. But that is just a reformulation of the original unfounded claim and obviously only begs the question. The somewhat complicated formulation of the MAR/CAR condition is not a pretentious though vacuous reformulation of this unjustified claim. The main point of the "simple device" lies in the use of conditional probabilities that provide the connection between the distribution of Y and that of Y^*. In particular, the first version of the MAR (or CAR) condition is a statement about the probability of the type of response that is chosen given the underlying state of the respondent. This is what Little and Rubin call a "missing data mechanism". While it can only be formulated within a model for the joined distribution of (Y, Y^*), it certainly provides a prescription to simulate incomplete data from any assumption on the distribution of the variable of interest, Y. This model construction permits to argue about the form of the "mechanism", about the validity of the MAR or CAR condition, and about consequences of deviations from it. And it is precisely this feature that makes the "simple device" potentially useful.

Still, the special formulation used here might look unfamiliar. Most texts, including the well known text book by Little and Rubin, formulate the MAR/CAR condition using a further random variable indicating whether data are missing or not. However, using subsets of $\mathcal{Y}$ to present incomplete data opens the way to discuss many types of incomplete data including truncated, censored, grouped, or heaped data in a unified way. Compared to pure missing data problems these general types of incomplete data provide both more information and more challenging problems. The latter aspect enriches the discussion of the appropriateness of the "simple device" in general and of certain formulations in particular.

The missing at random condition can be formulated only within the probability model. It has no real world counterpart. While the random variable Y^* of the model may be identified with the observed values of $Y^*(u), u \in \mathcal{U}$, the empirical counterpart $Y(u), u \in \mathcal{U}$ of the random variable Y is only partially known. Therefore, the missing at random conditions cannot be translated into conditions on empirical distributions. No reformulation of the missing at random conditions will help in this respect.[4]

Moreover, there is no way to formulate a "missing data mechanism" within the sampling theory framework. Using its bookkeeping variables allows to express assumptions pertaining to the real world such as: The percentage of employed among the non-respondents is 80%. It is plain that it might have been otherwise. But there is no way to argue for a particular value. In the sampling approach values of Y and Y^* are taken to be fixed for each interviewee. There is no formal way to express the idea of a tendency to answer to an interview request. On the contrary, the approach explicitly rules out any speculations about other possible responses and employment statuses than those that actually obtain. But such speculations are the essence of the "simple device".

It is certainly not enough simply to invoke the missing at random "assumption" as an assumption about the social world and then proceed using an estimate that simply ignores missing or incomplete observations. It is apparent that neither the "simple device" nor the MAR condition as such justify the use of a point estimate of 78% for the employment rate. Nor does it justify the general tendency to ignore missing values and incomplete data and to use only complete data.

The original formulation of the missing data problem concerned the percentage of employed among a certain group and the problem encountered was that not all people answered to the respective question. In this

[4] A more formal statement including necessary regularity conditions are discussed in the next Chapter.

setting, giving a particular percentage as the result of the survey can only be justified by appeal to convention and convenience or to experiences gained in more favourable circumstances. The "simple device" redefines this simple problem by first introducing a lot of mathematical structure and then solving a mathematical problem, not the one the survey statistician or the sociologist is interested in. The simple bookkeeping quantities (Y, Y^*) are transformed into random variables. Then the problem of missing values is translated into one of formulating a stochastic relation between Y and Y^*, a relation that has no direct empirical counterpart. And finally, sets of "assumptions" are produced that would guarantee the identification of the distribution of Y from the distribution of Y^*.

Before one may enjoy the merits or otherwise of the approach, one should therefore answer to the objections of its critics. The main objection, at least from the perspective of the survey statistician, is that the probability model and in particular the stochastic relation between Y and Y^* introduces additional mathematical structure far beyond the usual practises. But introducing additional structure may simply be a form of begging the question presupposing answers that otherwise would not exist. Moreover, in so far as some of the additional concepts are open to empirical scrutiny they may fail to be useful or applicable. Dalenius' and Morgenstern's remarks, if correct, would indicate that the prime motivating areas of applications of the statistical approach are exactly those where it lacks applicability.

The second objection pertains to the missing at random condition. Taken as an assumption about the real world, the condition is generally said to be untestable. In fact, in the example $Y(u) \in Y^*(u)$ by construction and the same relation is used in the probability model. Since only $Y^*(u)$ is empirically accessible any joint distribution of $Y(u, .)$ and $Y^*(u, .)$ that obeys the restriction is compatible with the data. And there is nothing else in the social world that corresponds to the MAR/CAR condition. It would be strange or at best elliptical to say that the missing at random assumption would hold for the interviewees of the National Longitudinal Survey of Youth. After all, it is certainly not a property or characteristic of the respondents. Nor does it pertain to the sample drawn. Neither is there a direct connection between the social world and probability statements as used in the “simple device”. Tryfos formulated:

> To our knowledge, there is no randomizer consulting a table of random numbers to determine, say, the value of an unknown determining factor. We know of no gambling “Nature”, no “invisible hand”, divine or otherwise, rolling dice or drawing from an urn. (Tryfos 2004: 69)

Nor is it possible to “apply” the model to the social world. Matheron clearly stated that

> ...the notion of probability is of a mathematical, not empirical nature, and thus the notion of objectivity as it is accepted in the positive sciences is not relevant to it. ... [N]obody has ever applied either the theory of probability or for that matter any other mathematical theory, to reality. One can only “apply” to reality real (physical, technical, etc.) operations, not mathematical operations. ... In other words, it is always to *probabilistic models*, and only to them, that we apply the theory of probability. (Matheron 1989: 27)

Nor does the probability model “describe” the world. Or at least, models are nearly never used to describe the world. Rather, descriptions in science document observations and facts. Statements in descriptions relate to facts. Statements made within a probabilistic framework—in particular the MAR and CAR conditions—do not.

Even though some of the proponents of the “simple device” and many practitioners deliberately ignore the distinction, the critics were well aware of it. In fact, Dalenius was speaking of “assumptions of unknown validity about probabilities”, thus referring to the probability model and not to the data or social reality. Within a given probability model, what then is the validity of assumptions of the MAR/CAR variety? It might seem to be an arbitrary mathematical prerequisite similar to the introduction of the probability axioms. But then there would be no point in discussing it, not even a point for calling it an assumption. One does not say “1 + 2 = 2 + 1 because I assume addition of natural numbers to be commutative”. Perhaps it is best seen as a modelling decision, a constitutive decision that “defines the general framework within which we shall operate and determines the choice of the tools we use" (Matheron 1989: 52). But such an interpretation is reasonable only for the introduction of the basic probabilistic framework for (Y, Y^*). That choice determines the tools we use. The MAR/CAR conditions do not determine the general framework but define a very special case of the general model.

On the other hand, the MAR/CAR conditions are not open to empirical assessment. In this respect, they are similar to the assumption of stochastic independence that is invoked in many social science models. While the latter assumption is not empirically accessible either, it does not prejudice a particular solution to a problem but sets the frame of reference in which solutions are formulated. The MAR/CAR assumptions are different in that they do lead to a particular solution. And this solution may turn out to be unreasonable, even though a decision can not be based on the available data. For such conditions, criteria for their appropriateness and usefulness are required beyond the consistency within a given model. Except for other data sources or general experience, the only possible base for a critical assessment of the MAR/CAR condition must rely on an operational model, a prescription on how to simulate the incomplete data within the model that avoids to exploit information on the complete data model above that revealed by the incomplete data. Procedural models that can be simulated without presupposing a knowledge of the "truth" have been developed for certain non-MAR/CAR models as well. An assessment of the merits of MAR/CAR or alternative conditions may then be based on a comparison of the outcomes of procedurally specified sub-models. Criteria for such comparisons, and for the appropriateness of conditions that are neither open to empirical scrutiny nor are constitutive for the model construction, have only recently been proposed.

I will discuss these problems in some detail in the first, newly written chapter. It presents a general probabilistic formulation of the incomplete data problem and the corresponding solution sets. I then proceed to study several alternative mathematical representations of the solution sets which will help to decide whether the probabilistic framework adds possibly too much structure, thus answering the first concern of the critics. In a second step, several formulations of the MAR/CAR conditions are introduced. Differences between these formulations have given rise to many confusions. A closer look at the differences, however, reveals that they are closely connected to the structure of incomplete data as represented in probabilistic models. One particular formulation, the one that relies on the introduction of additional random variables (e.g. indicators of missing values) to express MAR/CAR conditions, will in general introduce further restrictive elements into the analysis that go far beyond the introduction of a probability model per se. Such constructions are very useful since they allow to formulate MAR/CAR conditions even in cases that can not be couched in terms of a joined distributions of some (Y, Y^*). And they provide a simple language to state MAR/CAR conditions in terms of these variables. On the other hand, it turns out that such formulations are ambiguous: There are in general several versions, some of which will declare (Y, Y^*) to be MAR/CAR while others will contradict. I will conclude that such additional random elements will obfuscate the analysis as long as there is a reasonable mathematical structure of $(\mathcal{Y}, \mathcal{Y}^*)$. In a final step, various procedural models for the MAR and CAR conditions are introduced. These serve to study the special status of the MAR/CAR conditions in models of incomplete data. The chapter closes with some remarks on criteria of the appropriateness of the MAR/CAR conditions.

The next chapter, Chapter 3, is also written for the present document. It complements the previous chapter, discussing a practical problem using the tools developed so far. The problem is that of inferring the distribution of life lengths of a parent generation from the information provided by their children in a survey. The survey participants were asked for the birth dates of their parents and, if they died before the survey date, the date of death. Such a data structure combines truncated and censored data with a specific selection of the observations. Information on life length is obviously only available for those who had children so that deaths in infancy are never reported. Parents who are still alive at the time of the survey contribute censored observations. And parents with several children are possibly over-represented in a survey. The interplay of several forms of incompleteness makes this case study particularly valuable in demonstrating several versions of the “simple device” and their consequences.

Chapter 4 takes up the general discussion, this time concentrating on the selectivity of observations. In the simplest case, selection models are just a variant of missing data models in that information on a certain group of people is unavailable. But selectivity becomes a much more interesting problem when partially complete data are considered along with the completely missing case. A further extension of the basic model discussed in Chapter 2 is the introduction of conditional incomplete data models. Much more challenging variants of incomplete data models are models of self-selection. They are prevalent in many parts of micro-economics and sociology. The Roy model and the particular case of Heckman's model are but two examples that fit in the general framework. Such models clearly violate the CAR condition but still permit to deduce a unique answer from observational data, at least when some additional "assumptions" are invoked. One of the questions that prompted the article was whether arguments within the probability model would justify such non-CAR models. Another question that is investigated is the stability of answers from CAR or non-CAR models to minor variations in assumptions on the selection "mechanism". The newly written appendix provides some hints especially on the latter question. Techniques of sensitivity analyses have seen some important developments since the article was published. While the article surveys the main statistical methods suggested to deal with incomplete and self-selected data, the appendix updates hints to this rapidly growing literature. It also contains a short discussion on the appropriate standards by which to judge statistical methods via asymptotic considerations.

The next chapter, Chapter 5, discusses statistical approaches to causal inference. The predominant approach in statistics nowadays is based directly on the missing data paradigm. It defines a cause, or, perhaps better, the causal effect of the value of the "variable" $X(u) = x$ on $Y(u)$ to be the difference between $Y(u) \mid X(u) = x$ and $Y(u) \mid X(u) \neq x$. This is a missing data problem since for any given unit $u \in \mathcal{U}$, only either $X(u) = x$ or $X(u) \neq x$ but not both are observed. In consequence, most statistical approaches rely on some form of the CAR condition to identify the "causal" effect of X. The article provides an alternative relying on the timing of events. The approach identifies events as the relata of a causal relation and is thus closer to the classical philosophical discussion than that based on variables. Also, since humans can bring about events and since sequences of such events constitute an important topic in the social sciences it is better suited for social science applications than a counterfactual account. And finally the approach avoids to mimic (statistical) experiments that inspire the counterfactual approach. The analogy with experiments that is sought by other approaches may in many cases of interest to the social scientists turn out to be counterproductive.

The new appendix to the chapter starts with a short review of the counterfactual approach as it is discussed in the recent statistical and social science literature. It then illuminates the direct connection between the approaches to incomplete data problems and statistical versions of counterfactual accounts of causality. Recent progress in the two main problem areas of the event based approach, the notion of autonomy and the definition of local independence, are also reviewed.

Chapters 6 and 7 deal with the special case of censored data. The first proposes a potentially very useful extension of non-parametric techniques for censored data using the mean residual life function. Its main technical contribution is a thorough investigation of variance estimators based on a representation of Kaplan-Meier integrals in terms of independent summands. The main contribution in the present context of incomplete data, however, is the reminder that some statistical methods within the incomplete data paradigm can be re-expressed in a way that both fit the standard statistical methods geared to independent observations and the requirements of incomplete data methods. Chapter 7, on the other hand, provides a discussion of some constraints on the form of hazard functions when a particular constraint, that of periodic effects, is invoked. Such constraints are often present in labour market studies as well as in marketing, and these examples provided the motivation of the article. The constraints imply restrictions on the form of the influence of covariates that can be seen as a restriction on the possible incomplete data models compatible with the constraints when covariates are present.

Chapter 8 discusses a regression model for multivariate censored observations. Historically, the non-parametric multivariate censored data model has been one of the inspiring problems for a general account of incomplete data. It is one of the examples where the non-parametric maximum likelihood estimator is inconsistent. One of the consistent alternatives deliberately coarsens the data even further by grouping both event and censoring times. Perhaps unexpectedly, incomplete data may help in stabilising estimation problems. The article exploits this fact for the construction of a locally efficient regression estimator. Its construction is based on the missing information principle, a reformulation and extension of the CAR condition that shows how scores and other likelihood quantities transform when some data are partially incomplete. The appendix reviews new developments in the estimation theory with incomplete data. Recent variations and extensions of the Buckley-James estimator suggested in the paper are also discussed.

The last Chapter 9 investigates the effect of incorrect model choice in regression models with linear predictors. It turns out that even when the model is mis-specified the relative size of regression coefficients is correctly estimated. This result, it is noted, is directly connected with a situation where the data for the dependent variable are only incompletely observed. Thus, a form of duality between incomplete data models and the analysis of mis-specified regression models is established. This type of partial duality, though probably known in special cases and for a long time, is certainly not yet explored to the same extent as other areas of incomplete data analysis. The new appendix lists a few more recent hints to a duality between incomplete data and wrong statistical models.

Finally, the work is supplemented by an appendix documenting an implementation of CAR compatible non-parametric distribution estimators within TDA (Transition Data Analysis, www.stat.ruhr-uni-bochum.de/tda.html), a computer package designed specifically to deal with censored and otherwise incomplete data. It is the base for Buckley-James type estimators discussed in Chapter 8 and can easily be used as a building block for regression problems with more general types of incomplete data.

2

Probabilistic Models of Incomplete Data

While the treatment of missing data has a long tradition in statistics, it was only 30 years ago that Rubin proposed to study such problems systematically from a probabilistic point of view. In retrospect, Rubin and Little dubbed the introduction of random variables, probability distributions and expectations to the study of missing data a "simple device". This chapter studies the ramifications of the "simple device" in the context of general incomplete data. It outlines a theory of probabilistic incomplete data models that is general enough to cover all data structures discussed in later chapters. And it provides the basic framework used in the statistical approach to incomplete data. Within this framework, some of the problems mentioned in the introduction can be answered. Indeed, the very introduction of a probability model may at first sight seem extraneous to the problem of analysing incomplete data. The use of probability theory may simply presuppose solutions that otherwise do not exist. It is not clear at the outset whether the introduction of the "simple device" will prejudice its solutions, thus just begging the original question.

I will show that this is not the case in the following section. There I also explore the mathematical structure of incomplete data models and relate them to other well known structures in mathematics. The second section introduces several versions of the coarsening at random (CAR) model. It turns out that the differences between CAR formulations are closely connected to the structure of incomplete data which therefore is an essential part of the "simple device" of Rubin and Little. The discussion will clarify several misunderstandings about the role of CAR models that pervade the current literature. Among these are the empirical content of CAR models and the uniqueness of CAR models. The next sections considers the role of missing indicators that play such a prominent role in the now classical version of the "simple device". It is shown that the introduction of the missing indicator (or of any other further random variables like censoring variables) extends the field of applications considerably. But it does so at the cost of making the definition of CAR dependent on modelling decisions that can not be justified from considerations of the structure of incomplete data alone. In this sense, the introduction of random variables indicating the extent of incompleteness goes beyond a probabilistic re-expression of the incomplete data problem. A last section concludes by taking up the discussion of criteria of applicability of probabilistic incomplete data models.

2.1. The Structure of Incomplete Data

Statistical data as well as random variables are commonly defined as functions with values in a previously defined space $\mathcal{Y}$. Statistical variables take as their domains of definition a list representing the target population, say $\mathcal{U}$. On the other hand, random variables are defined on a probability space Ω equipped both with a probability measure Pr(.) and a set of subsets of Ω to which probabilities can be assigned. The device of random variables as functions on Ω serves to induce further probability distributions and to define operations between random variables. In contrast, statistical variables are a device to keep track of properties of the members of the population. In both cases, however, the range space is chosen by the researcher. It is an a priori decision in Matheron's sense, constituting the questions that are contemplated.

More often than not, the detailed information necessary for the assessment of a statistical variable is not available. Employment status may be thoroughly defined, say following the ILO standard. Thus the range of a corresponding variable is well defined. But if people are asked about their employment status using that standard, it may not fit too well with the knowledge of the respondents. In consequence, respondents may only provide a rough characterisation of their employment status or even do not answer at all.

This situation can be represented by a further set $\mathcal{Y}^*$, a set of subsets of $\mathcal{Y}$. As an example, suppose that employment status is differentiated into the three categories 'employed', 'unemployed' and 'out of labour force' coded consecutively as $\mathcal{Y} = \{1, 2, 3\}$. Suppose that respondents are reluctant or unable to differentiate between unemployment and out of labour force. Further they may refuse to answer at all. Then the observed data can be represented by

$$\mathcal{Y}^* = \{\{1\}, \{2\}, \{3\}, \{2, 3\}, \{1, 2, 3\}\} \subseteq \mathcal{P}(\mathcal{Y}) \setminus \{\emptyset\}$$

where $\mathcal{P}(\mathcal{Y})$ is the power set of $\mathcal{Y}$. Here a non-response is represented by $\{1, 2, 3\}$ while someone who is unwilling to differentiate between unemployment and being out of labour force is represented by $\{2, 3\}$. Exact answers are represented by $\{1\}$, $\{2\}$, or $\{3\}$.

In general, let

$$Y : \mathcal{U} \longrightarrow \mathcal{Y}$$

be a statistical variable. Then coarse data can be represented by a statistical variable Y^*

$$Y^* : \mathcal{U} \longrightarrow \mathcal{Y}^* \subseteq \mathcal{P}(\mathcal{Y}) \setminus \{\emptyset\}$$

such that for all $u \in \mathcal{U}$: $Y(u) \in Y^*(u)$. The last condition excludes the possibility that respondents are misrepresenting their situation. I will term this requirement the *consistency condition*. In a purely set based model of incomplete data such a restriction is necessary since otherwise there would be no connection between Y and Y^* at all.

The consistency condition rules out some situations that sometimes are considered as incomplete data problems. If all respondents possibly mis-classified their situation by indicating another employment than the one they are in, then the consistency condition would force the choice of $\{1, 2, 3\}$ as a representation of all answers. But that would be completely uninformative.

Fortunately, most incomplete data problems encountered in social science applications, in particular grouped, truncated, and censored data, do have a representation in terms of subsets. Consider grouped data: If $\mathcal{Y}$ is an ordered set represented by $\{0, 1, \dots, \tau\}$, and if $[y_i, y_{i+1}]$ are intervals into which some observations are grouped, then $\mathcal{Y}^* = \{\{0\}, \{1\}, \dots, \{\tau\}, [y_0, y_1], \dots, [y_{k-1}, y_k]\}$ may be used to represent the situation. Such a grouping together with the possibility to provide "exact" values is often employed in survey questions. An example from the ALLBUS is discussed in Chapter 4. A similar case of the grouping together of observations that is not restricted to intervals is given by a general function $g\colon \mathcal{Y} \to \mathcal{Z}$ that can be represented as $\mathcal{Y}^* = \{g^{-1}(\{z\}) \mid z \in \mathcal{Z}\}$.

Another important case of incompleteness is provided by censored or truncated data. If $\mathcal{Y}$ is again the ordered set $\{0, 1, \dots, \tau\}$, then censored observations, i.e. observations for which it is only known that an event happened after a certain period, can be represented by $\{y, \dots, \tau\}$ such that $\mathcal{Y}^*$ becomes $\{\{0\}, \{1\}, \dots, \{\tau\}, \{0, \dots, \tau\}, \{1, \dots, \tau\}, \dots.\}$. This structure is studied in most of the later chapters.

A parallel construction can be used within a probability model without any change, resulting in

$$(Y, Y^*)\colon \Omega \longrightarrow \mathcal{Y} \times \mathcal{Y}^* \tag{2.1}$$

such that

$$Y(\omega) \in Y^*(\omega) \text{ for all } \omega \in \Omega \tag{2.2}$$

This will also be called *consistency condition.*[1]

[1] The condition (2.2) can be weakened to $\Pr(Y \in Y^*) = 1$. This might be necessary if Ω is chosen uncountable, e.g. to model continuous covariates. But with finite sets $\mathcal{Y}$ as used here, and without the necessity to model additional information, nothing essential is gained since one can always choose a finite Ω with $\Pr(\{\omega\}) > 0$ for all ω.

One aspect of incomplete data can be studied by just considering the structure of $\mathcal{Y}^*$. To facilitate the discussion, I introduce three simple examples that will be considered throughout this Chapter.

Example 1
Suppose again that employment status is differentiated into the three categories 'employed', 'unemployed' and 'out of labour force' coded consecutively as $\mathcal{Y} = \{1, 2, 3\}$. Suppose further that respondents are reluctant or unable to differentiate between unemployment and out of labour force. Further they may of course refuse to answer at all. Then the observed data can be represented by

$$\mathcal{Y}_1^* = \{\{1\}, \{2\}, \{3\}, \{2, 3\}, \{1, 2, 3\}\}$$

Example 2
Respondents in addition to the previous reporting conditions would sometimes only report whether they are employed or unemployed vs. being out of labour force. Then the support of Y^* becomes

$$\mathcal{Y}_2^* = \{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{2, 3\}, \{1, 2, 3\}\}$$

Example 3
The second example can be changed slightly by deleting $\{3\}$. Thus nobody would report to be out of labour force. The resulting $\mathcal{Y}_3^*$ becomes

$$\mathcal{Y}_3^* = \{\{1\}, \{2\}, \{1, 2\}, \{2, 3\}, \{1, 2, 3\}\}$$

The sets of coarsened data are partially ordered by set inclusion. Thus, in the first example, one has $\{2\} \prec \{2, 3\} \prec \{1, 2, 3\}$. And in the second example $\{2\} \prec \{1, 2\}, \{2, 3\} \prec \{1, 2, 3\}$. The ordering relations of the elements of the three examples can be read from the corresponding Hasse diagrams presented in Figure 2.1

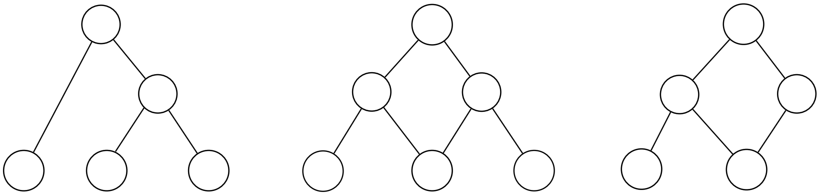


Figure 2.1.: Hasse diagrams for the sets $\mathcal{Y}_1^*$, $\mathcal{Y}_2^*$ and $\mathcal{Y}_3^*$. The top nodes correspond to the set $\{1, 2, 3\}$, the bottom nodes to the sets $\{1\}, \{2\}, \{3\}$.

2.1.1. A Typology of Missing Data Structures

Example 1 is special in that the elements of $\mathcal{Y}_1^*$ are *hierarchically ordered.* By this I mean that for all elements $y_1^*, y_2^* \in \mathcal{Y}^*$ we have either $y_1^* \subseteq y_2^*$ or $y_2^* \subseteq y_1^*$ or $y_1^* \cap y_2^* = \emptyset$. Another way to describe this special situation that will be of importance later on is that $\mathcal{Y}^*$ is the union of a sequence of refinements of a partition of $\mathcal{Y}$. Here, the sequence of partitions is $\{\{1, 2, 3\}\}, \{\{1\}, \{2, 3\}\}, \{\{1\}, \{2\}, \{3\}\}$.

The case of hierarchically ordered incomplete data structures is encountered in many common models of incomplete data situations. The most obvious one is when only completely missing or exact observations of a single variable are considered. Then the resulting $\mathcal{Y}^*$ is trivially hierarchically ordered.

If observations are discrete and possibly right censored durations, censored observations are presented by sets of the form $\{y, \dots, \tau\}$. Such observations just tell that an event happened after period $y - 1$, but they don't reveal the exact time of the event. Rearranging terms, the set $\mathcal{Y}^*$ becomes $\{\{0, \dots, \tau\}\} \cup \{\{0\}, \{1, \dots, \tau\}\} \cup \{\{0\}, \{1\}, \{2, \dots, \tau\}\} \cup \dots$ so that it is a union of refinements of a partition and thus hierarchically ordered.

Grouped data provide another important example. If $\mathcal{Y} = \{0, \dots, \tau\}$ and if there is just one level of grouping $\mathcal{Y}$ into disjoint intervals $[y_i, y_{i+1}]$ whose union is $\mathcal{Y}$, then the $\mathcal{Y}^*$ resulting from taking the union of the intervals with the set of all singletons and possibly the whole set (completely missing data) will result in an hierarchically ordered $\mathcal{Y}^*$.

Adding further levels of grouping still will lead to hierarchically ordered models as long as the further levels are either refinements or coarsenings of the original grouping.

Perhaps the most important example is provided by *monotone missing data* patterns. Consider several variables $Y_1, \ldots, Y_k$ with values in $\mathcal{Y}_1 \times \ldots \times \mathcal{Y}_k$. Suppose that each variable is either observed or completely unobserved. If the pattern of missingness is such that there is an ordering of the variables, say $(1) < (2) < \ldots < (k)$, and such that if $Y_{(i)}$ is missing (i.e. $Y^*_{(i)} = \mathcal{Y}_{(i)}$) then all $Y_{(j)}$ are missing as well as long as $(i) < (j)$, the pattern is called a monotone missing pattern (see e.g. Little, Rubin 2002: Chap. 7.4). A typical element of the set $\mathcal{Y}^*$ is

$$\{(y_{(1)}, \ldots, y_{(i-1)})\} \times \mathcal{Y}_{(i)} \times \ldots \times \mathcal{Y}_{(k)}$$

and the union of these sets over all tuples $(y_{(1)}, \ldots, y_{(i-1)})$ forms a partition of $\mathcal{Y}_{(1)} \times \ldots \times \mathcal{Y}_{(k)}$. Furthermore, if $\mathcal{P}_i$ is such a partition with smallest index (i) such that all the values of $Y_{(j)}$ with index (j) larger than (i) are missing, the partition $\mathcal{P}_j$ with $(j) > (i)$ is a refinement of $\mathcal{P}_i$. Thus, monotone missing data patterns are hierarchically ordered.

In section 2.2, I will show that the CAR condition for hierarchically ordered models can be treated very much like a single variable which is either exactly observed or completely missing. But how can one justify the choice of such a simple model? In the case of monotone missingness the order of the variables may arise from an ordering of observations according to the time they were made. This is often the case in panel studies. Moreover, in that case, monotone missingness just excludes the possibility of returning to the panel population after some temporary withdrawal. This restriction is often part of the design of a panel study. In other cases, it might be seen as a reasonable approximation to the design.

But in many other models such an assumption seems to be rather far fetched. Consider regression models with several covariates where some or even all of the covariates are incompletely observed. That situation can be formalised quite similarly to the case of monotone missingness. Nevertheless, a regression context will nearly never allow to justify a particular order of incompleteness information. And even if one is thought to be adequate, a few data may prove the assumption to be inadequate. Or, arguing the other way around, even if a given data set is consistent with a certain order of incompleteness, this may not be expected in general and not even for similar data sets. Still, it would be inadequate to decide on the choice of the model for the incompleteness structure solely based on one (or several) data sets. It is often much easier to envisage broader forms of incompleteness than those actually present in a data set since it may then be possible to largely reduce the modelling burden. The choice of the incompleteness structure will therefore always contain preliminary modelling decisions that are informed by data aspects but are not determined by it.

Since hierarchically ordered incomplete data structures and their prototype of missing values in single variables share many nice properties, the study of the value of the "simple device" might in fact be prejudiced when only these special cases are considered. A less demanding data structure is given by a union of several partitions that need neither be refinements nor coarsenings of each other. An example in point would be data grouped into several types of categories which may overlap. Example 2 provides another example. $\mathcal{Y}_2^*$ is not hierarchically ordered since $\{1,2\} \cap \{2,3\} \neq \emptyset$ but neither $\{1,2\} \subseteq \{2,3\}$ nor $\{2,3\} \subseteq \{1,2\}$. But it is the union of several partitions, since one may write $\mathcal{Y}_2^* = \{\{1,2,3\}\} \cup \{\{1\},\{2,3\}\} \cup \{\{3\},\{1,2\}\} \cup \{\{1\},\{2\},\{3\}\}$. Another prominent example is given by current status data where for a given point in time it is only known whether an event did occur previously or not. Then the incomplete data of event times are of the form $\{0,\ldots,y\}$ if the event happened before time y. Or they take the form $\{y+1,\ldots,\tau\}$ when the event did not yet happen at inspection time. Clearly, this forms a partition of $\mathcal{Y}$ for each inspection time y and thus $\mathcal{Y}^*$ is a union of partitions. But these partitions are not refinements of each other.

The third example, $\mathcal{Y}_3^*$, is neither hierarchically ordered nor even the union of partitions of $\mathcal{Y}$ since for the element $\{2\}$ the only disjoint set in $\mathcal{Y}_3^*$ is $\{1\}$ but there union is not all of $\mathcal{Y}$. The consideration of such examples will provide the background against which several formulations of CAR can be compared. It will transpire that modelling decisions pertaining only to the choice of $(\mathcal{Y}, \mathcal{Y}^*)$ have important consequences for the evaluation of the "simple device".

2.1.2. Graphical Representations of Coarsened Data

Depicting the order of the elements of $\mathcal{Y}^*$ will as well exhibit which elements of $\mathcal{Y}$ belong to which sets in $\mathcal{Y}^*$ if the latter contains all one-element sets. If this is not the case, one would need to indicate this additional information, e.g. by providing the incidence matrix with rows presenting the elements of $\mathcal{Y}$ and columns the elements of $\mathcal{Y}^*$. In the matrix, the (i, j)-th element is 1 if the i-th element of $\mathcal{Y}$ is an element of the j-th element of $\mathcal{Y}^*$, 0 otherwise. In the first example one gets

	$\{1\}$	$\{2\}$	$\{3\}$	$\{2,3\}$	$\{1,2,3\}$
1	1	0	0	0	1
2	0	1	0	1	1
3	0	0	1	1	1

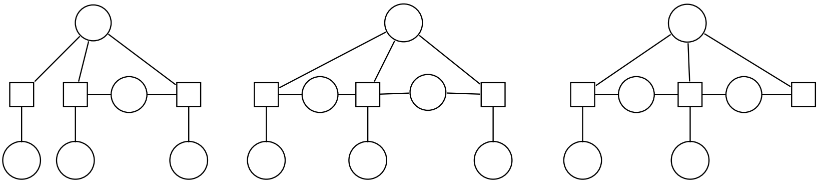


Figure 2.2.: Levi graph for the sets $\mathcal{Y}_1^*$, $\mathcal{Y}_2^*$ and $\mathcal{Y}_3^*$. The top circles correspond to the set $\{1, 2, 3\}$, the bottom circles to the sets $\{1\}, \{2\}, \{3\}$. The squares correspond to the elements of $\mathcal{Y}$, 1,2,3.

This could be visualised by a hypergraph whose vertices are the elements of $\mathcal{Y}$ and the edges indicate the composition of the elements of $\mathcal{Y}^*$. An equivalent representation is the Levi graph, a bipartite graph with the first set of vertices being the elements of $\mathcal{Y}$ and the second set being the elements of $\mathcal{Y}^*$. There is an edge between vertices of the two sets iff $y \in y^*$. The Levi graphs of the three examples are given in Figure 2.2.

The Levi graph and the hypergraph of an incidence matrix are obviously in one-to-one correspondence.

Jaeger (2005b) argued that it is better to use the dual of the hypergraph where the elements of $\mathcal{Y}^*$ are the vertices and the edges indicate which elements of $\mathcal{Y}$ belong to them. The resulting hypergraphs for the three examples are depicted in Figure 2.3.

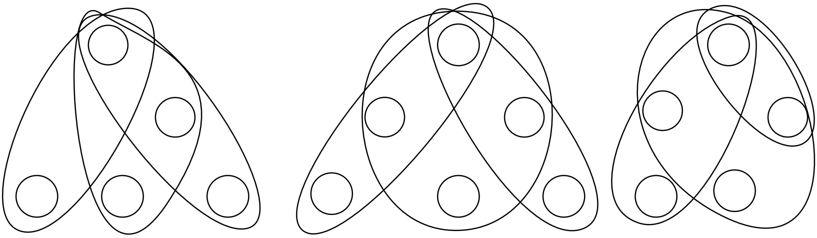


Figure 2.3.: Hypergraphs for the sets $\mathcal{Y}_1^*$, $\mathcal{Y}_2^*$ and $\mathcal{Y}_3^*$. The layout is the same as in the previous Figure, the top nodes correspond to the set $\{1, 2, 3\}$, the bottom nodes to the sets $\{1\}$, $\{2\}$, $\{3\}$. The edges indicate the elements of $\mathcal{Y}$.

Taking an extreme example one may consider

$$\mathcal{Y}_4^* \coloneqq \{\{1, 2\}, \{2, 3\}\}$$

so that the only information available would be whether someone was employed or unemployed, or unemployed or out of labour force. There is no ordering between the two sets. There are no singletons either and so the information on which element of $\mathcal{Y}$ belongs in which of the two coarsened sets in $\mathcal{Y}^*$ must be added. The corresponding hypergraph is given in Figure 2.4.

In this hypergraph, the edge representing 2 (the outer ellipse) contains two proper subsets, the edges representing 1 and 3 (the inner circles around the two circles representing the sets $\{1, 2\}$ and $\{2, 3\}$). Jaeger (2005b) argued that whenever the hypergraph contains properly nested edges then the CAR condition can not hold (except for degenerate cases). The non-existence of CAR models would then follow just from the

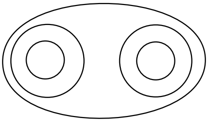


Figure 2.4.: Hypergraph for the set $\mathcal{Y}_4^*$. The innermost circles represent the sets $\{1, 2\}$ and $\{2, 3\}$. The outer circles represent the elements 1 and 3, and the out-most oval represents the element 2.

structure of the incomplete data and without considering any more details like the distribution of Y^*.

There is an interesting connection of this example with the famous Monty Hall problem: A contestant at the Monty Hall show has to choose between three doors 1, 2, 3, behind one of which there is a price. Behind the others there is a goat. If the contestant has chosen door 2, say, the host opens one of the other doors which reveals a goat. The information that the contestant can expect after his choice of the door is that the price is behind either the doors $\{1, 2\}$ or $\{2, 3\}$. In particular, the Monty Hall problem has the structure given in Figure 2.4. Next, the host reveals which of the two sets from $\mathcal{Y}_4^*$ obtain. Given this information, the candidate can either change his choice of a door or stick with the original one. But this opening of a door by the host is informative for the location of the price: Informally, if the rules of the game were uninformative about the location of the price, the opening of, say, door 3 by the host should only reveal that the price is either behind door 2 or behind door 1 ($Y^* = \{1, 2\}$). Put differently, the decision to reveal $\{1, 2\}$ ought not depend on the location of the price. Now if the price is behind door 1, the rules of the game force the host to open door 3. But then an uninformative rule would force him to open door 3 also when the price is behind door 2. By the same argument, the host must open door 1 (revealing $Y^* = \{2, 3\}$) when the price is behind door 3. An uninformative rule would lead him to open door 1 also when the price is behind door 2. In consequence, if an uninformative rule existed it would lead to inconsistent behaviour (opening both door 1 and door 3 when the price is behind door 2). Thus there is no uninformative rule and the CAR condition certainly does not hold.[2] This example provides a connection between the incomplete data literature and that on probability dynamics in the Bayesian literature.[3]

2.1.3. Distributions Consistent with Coarsened Data

The last piece of information required for the representation of incomplete data is a distribution on $\mathcal{Y}^*$. This might either be the distribution of a statistical variable Y^* representing the observed frequencies of incomplete data. Or, as in the Monty Hall example, it might be a probability distribution representing the "rules of the game". The task is to infer something useful about the distribution of Y on $\mathcal{Y}$ from a distribution on $\mathcal{Y}^*$.

In any case, instead of taking a particular subset $\mathcal{Y}^*$ of the power set of $\mathcal{Y}$ as the support of the distribution, one might as well take the full power set and use zero probability assignments for certain elements of the power set to define the structure of incompleteness.[4] However, I prefer to separate the determination of a distribution of Y^* from the determination of its support encoded by a subset of the full power set, namely $\mathcal{Y}^*$. Often the distribution of Y^* is determined by the empirical distribution. But the structure of $\mathcal{Y}^*$ is only partly determined by the data. It might be chosen much larger than the data suggest, so that a particularly simple structure emerges. Or it might, as in the Monty Hall example, encode the rules of the game. There are also technical problems:

[2] This rather informal discussion presupposes that the price may be behind any of the doors with positive probability in order to derive a contradiction. If the price is always behind door 1 (and the contestant has chosen door 2 preliminarily), then there is a unique strategy of the host that is 'uninformative' in the above sense: He is forced to reveal $\{1, 2\}$. Similarly, if the price is always behind door 2, the host is free to either reveal $\{1, 2\}$ or $\{2, 3\}$. This will always be 'uninformative' whatever his strategy to choose between$\{1, 2\}$ and $\{2, 3\}$. These 'uninformative' possibilities are only ruled out by a further elaboration of the rules of the game.

[3] A short and vivid discussion is given by Jeffrey (2004: Chap. 3). The notion of CAR in probability dynamics is further discussed by Jaeger (2005b), de Cooman and Zaffalon (2003, 2004), Grünwald and Halpern (2003).

[4] The sets $y^* \in \mathcal{P}(\mathcal{Y})$ are called *focal elements* by Shafer (1976: 40). His notion of *core* as the union of all focal elements is a subset of my $\mathcal{Y}^*$.

With infinite sets $\mathcal{Y}$, it is in general impossible to assign probabilities to all subsets of $\mathcal{Y}$. In such applications, subsets of the power set must in any case be determined before a probability distribution can be assigned. Consequently, in the following a particular subset $\mathcal{Y}^*$ of the power set of $\mathcal{Y}$ is used as the support of the distribution of Y^*. The distribution may, however, assign zero probabilities to some further elements of $\mathcal{Y}^*$. On the other hand, it is much easier to work with probabilities defined on all of $\mathcal{P}(\mathcal{Y})$ so that I will without further mentioning use the extension of a probability on $\mathcal{Y}^*$ by stipulating $\Pr(Y^* = A) = 0$ for all $A \subseteq \mathcal{Y}$ such that $A \notin \mathcal{Y}^*$.

The distribution with support $\mathcal{Y}^*$ together with the consistency condition will imply restrictions on the distribution of the corresponding variable Y. Since $Y(.) \in Y^*(.)$ for all arguments of the variables, $\{Y^* = y^*\} \subseteq \{Y \in y^*\}$ for all $y^* \in \mathcal{Y}^*$ because if $Y^*(\omega) = y^*$, then necessarily $Y(\omega) \in y^*$ for (almost) all $\omega \in \Omega$. In particular,

$$\Pr(Y^* = y^*) \leq \Pr(Y \in y^*) \tag{2.3}$$

for all $y^* \in \mathcal{Y}^*$. A useful consequence is

$$\Pr(Y^* = y^*) = \Pr(Y^* = y^*, Y \in y^*) \tag{2.4}$$

Since also $\{Y^* = y^{*\prime}\} \subseteq \{Y \in y^*\}$ for all subsets $y^{*\prime} \subseteq y^*$, it follows that $\cup_{y^{*\prime} \mid y^{*\prime} \subseteq y^*} \{Y^* = y^{*\prime}\} \subseteq \{Y \in y^*\}$ and the inequality can be strengthened to

$$\sum_{y^{*\prime} \subseteq y^*} \Pr(Y^* = y^{*\prime}) = \Pr(Y^* \subseteq y^*) \leq \Pr(Y \in y^*) \tag{2.5}$$

The same reasoning applies not only to y^* but to arbitrary subsets $A \subseteq \mathcal{Y}$. Thus one finally arrives at

$$\sum_{y^{*\prime} \subseteq A} \Pr(Y^* = y^{*\prime}) = \Pr(Y^* \subseteq A) \leq \Pr(Y \in A) \tag{2.6}$$

for all $A \subseteq \mathcal{Y}$.

On the other hand, $\{Y = y\} \subseteq \cup_{y^* \ni y}\{Y^* = y^*\}$ since if $Y(\omega) = y$ then necessarily all possible values of $Y^*(\omega)$ must contain y (the empty set is excluded from $\mathcal{Y}^*$) and therefore

$$\Pr(Y = y) \leq \sum_{y^* \ni y} \Pr(Y^* = y^*) = \Pr(Y^* \cap \{y\} \neq \emptyset) \tag{2.7}$$

for all $y \in \mathcal{Y}$. An immediate consequence is that

$$1 = \sum_{y^* \ni y} \Pr(Y^* = y^* \mid Y = y) \tag{2.8}$$

for all y with $\Pr(Y = y) > 0$ since the right hand side is just the probability of the set $\cup_{y^* \ni y}\{Y^* = y^*\}$ conditioned on its subset $\{Y = y\}$. The equality states that the support of the conditional probabilities given $\{Y = y\}$ is the subset of elements of $\mathcal{Y}^*$ that contain y.

The inequality can be generalised to give

$$\Pr(Y \in y^*) \leq \sum_{y^{*\prime} \cap y^* \neq \emptyset} \Pr(Y^* = y^{*\prime}) = \Pr(Y^* \cap y^* \neq \emptyset) \tag{2.9}$$

This is not, in general, a strengthening of (2.7). In the case of $\mathcal{Y}_4^*$, $\Pr(Y = 1) \leq \Pr(Y^* \in \{1, 2\})$ but trivially $\Pr(Y \in \{1, 2\}) \leq \Pr(Y^* = \{1, 2\}) + \Pr(Y^* = \{2, 3\}) = 1$. In hierarchical data structures, however, the upper bound of $\Pr(Y = y)$ stays the same when the event $\{Y = y\}$ is replaced by the possibly larger event $\{Y \in y^*\}$ when y^* is the unique element of the finest partition of $\mathcal{Y}$ that contains y. Thus in hierarchical incomplete data structures it is always sufficient to consider the pairs of inequalities (2.5) and (2.9).

Since there is no best upper bound on subsets of $\mathcal{Y}$ in general, one has to check that

$$\Pr(Y \in A) \leq \sum_{y^* \cap A \neq \emptyset} \Pr(Y^* = y^*) = \Pr(Y^* \cap A \neq \emptyset) \tag{2.10}$$

for all $\emptyset \neq A \subseteq \mathcal{Y}$. Combining this with the respective lower bounds, the full set of inequalities becomes:

$$\sum_{y^* \subseteq A} \Pr(Y^* = y^*) = \Pr(Y^* \subseteq A) \leq \Pr(Y \in A)$$

$$\begin{aligned} &\leq \Pr(Y^* \cap A \neq \emptyset) \qquad (2.11)\\ &= \sum_{y^* \cap A \neq \emptyset} \Pr(Y^* = y^*) \end{aligned}$$

for all $\emptyset \neq A \subseteq \mathcal{Y}$.

While the extension to all non-empty subsets of $\mathcal{Y}$ introduces many redundant restrictions, it allows for a concise and symmetrical formulation. Setting

$$F(A) \coloneqq \Pr(Y^* \subseteq A) = \sum_{y^* \subseteq A} \Pr(Y^* = y^*)$$

$$T(A) \coloneqq \Pr(Y^* \cap A \neq \emptyset) = \sum_{y^* \cap A \neq \emptyset} \Pr(Y^* = y^*)$$

the inequalities can be abbreviated to

$$F(A) \leq \Pr(Y \in A) \leq T(A) \qquad (2.12)$$

But $\{Y^* \subseteq A\} = \{Y^* \cap A^c = \emptyset\} = \{Y^* \cap A^c \neq \emptyset\}^c$. Thus $F(A) = 1 - T(A^c)$. In particular, the upper bound for the set A^c, $\Pr(Y \in A^c) \leq T(A^c)$, is equivalent to $\Pr(Y \in A) = 1 - \Pr(Y \in A^c) \geq 1 - \Pr(Y^* \cap A^c \neq \emptyset) = F(A)$, the lower bound for the set A. Since the inequalities are to hold for all non-empty $A \subseteq \mathcal{Y}$, the consistency conditions can be shortened to the one sided condition

$$F(A) \leq \Pr(Y \in A) \eqqcolon \Pi(A) \qquad \forall\, \emptyset \neq A \subseteq \mathcal{Y} \qquad (2.13)$$

One may think of $F(A)$ as a distribution function of Y^*: It is monotone increasing with $F(\emptyset) = 0$ (since $\emptyset$ is excluded from $\mathcal{Y}^*$ by definition, $\Pr(Y^* = \emptyset) = 0$) and $F(\mathcal{Y}) = 1$. Moreover, knowing $F(.)$, one gets back the probabilities of all $y^* \in \mathcal{Y}^*$ by[5]

$$\Pr(Y^* = y^*) = \sum_{y^{*\prime} \subseteq y^*} (\text{-}1)^{|y^* \setminus y^{*\prime}|} F(y^{*\prime}) \qquad (2.14)$$

[5] This is the Möbius function from the combinatorics of partially ordered sets. See Aigner (1997: Chap. 4) who stresses the similarity with differentiation and integration and who presents many further connexions.

This is similar to the formula determining the probability of rectangles from distribution functions in higher dimensions.

In fact, $F(.)$ enjoys a stronger monotonicity property. It is *2-monotone* in that for $A, B \subseteq \mathcal{Y}$

$$F(A \cup B) \geq F(A) + F(B) - F(A \cap B)$$

since

$$\begin{aligned} F(A \cup B) &= \Pr(Y^* \subseteq A \cup B) \geq \Pr(Y^* \subseteq A \ \vee \ Y^* \subseteq B) \\ &= F(A) + F(B) - F(A \cap B) \end{aligned}$$

It is even *infinitely monotone* since for any k and any subsets $A_1, A_2, \ldots, A_k$ of $\mathcal{Y}$

$$F\left(\bigcup_{i=1}^{k} A_i\right) \geq \sum_{\emptyset \neq I \subseteq \{1,2,\ldots,k\}} (\text{-}1)^{|I|+1} F\Big(\bigcap_{i \in I} A_i\Big)$$

It can be shown that all infinitely monotone set functions on $\mathcal{Y}$ with $F(\emptyset) = 0$ and $F(\mathcal{Y}) = 1$ arise from a probability distribution of a random variable Y^* taking values in $\mathcal{Y}^*$.[6]

This characterisation of $F(.)$ as a normed infinitely monotone set function connects the necessary consistency condition with many other concepts suggested to deal with incomplete information. In particular, such an $F(.)$ is a *belief function* in the sense of Shafer (1976). Furthermore, Strassen (1964) and Huber (1976) suggested to use either 2-monotone or infinitely monotone distribution functions to express limits of measurement precisions in statistics.[7]

[6] This result is known as Choquet's theorem. See Nguyen (2006: 38) for a simple proof in the case of finite $\mathcal{Y}$. The general case is thoroughly discussed in the classical monograph of Phelps (2001) which was first published in 1965. Molchanov (2005) provides an updated review.

[7] See Weichselberger (2001) and Walley (1991) for broad reviews. There is also an interesting connexion with game theory where the set of probability measures $\Pi(A) \geq F(A)$ is called the core of a game. Characterisations of the core are discussed by Chateauneuf/Jaffray (1989) from a combinatorial point of view. Wallner (2007)

2.1.4. The Geometry of Consistent Distributions

The consistency conditions in the form of (2.13) pertain only to either the distribution of the random variables (Y, Y^*) or their statistical counterparts. As such, they do not encode whether in fact $Y(.) \in Y^*(.)$ or not. They merely record distributional consequences of the consistency requirement $Y(.) \in Y^*(.)$. But even these weak necessary conditions provide rather strong information about the set of possible distributions of Y.

Looking again at the first example and writing $p_{\{1\}}, \ldots, p_{\{1,2,3\}}$ for the density of the variable Y^*, p_1, p_2, p_3 for the density of the variable Y, the (non-vacuous) consistency requirements become

$$\begin{aligned} p_{\{1\}} \leq p_1 &\leq p_{\{1\}} + p_{\{1,2,3\}} \\ p_{\{2\}} \leq p_2 &\leq p_{\{2\}} + p_{\{2,3\}} + p_{\{1,2,3\}} \\ p_{\{3\}} \leq p_3 &\leq p_{\{3\}} + p_{\{2,3\}} + p_{\{1,2,3\}} \\ p_{\{2\}} + p_{\{3\}} + p_{\{2,3\}} \leq p_2 + p_3 &\leq p_{\{2\}} + p_{\{3\}} + p_{\{2,3\}} + p_{\{1,2,3\}} \end{aligned} \tag{2.15}$$

The last restriction is generally non-redundant since its upper bound can be smaller than the upper bound available from adding together the upper bounds for p_2 and p_3. Similarly, the lower bound will be larger than the lower bound implied by the second and third line.

For the next two examples, the restrictions are

$$\begin{aligned} p_{\{1\}} \leq p_1 &\leq p_{\{1\}} + p_{\{1,2\}} + p_{\{1,2,3\}} \\ p_{\{2\}} \leq p_2 &\leq p_{\{2\}} + p_{\{1,2\}} + p_{\{2,3\}} + p_{\{1,2,3\}} \\ p_{\{3\}} \leq p_3 &\leq p_{\{3\}} + p_{\{2,3\}} + p_{\{1,2,3\}} \end{aligned}$$

provides a further connexion to interval probabilities where monotonicity need not hold.

The set function $T(.)$ is called *capacity functional* in the literature on random sets and *upper probability* in the literature on imprecise probabilities. It has complementary properties to the ones of the set function $F(.)$. Molchanov (2005: Chap. 1) provides a comprehensive review. Applied to singletons, $T(\{y\})$ is the *inclusion probability* of sampling theory.

$$\begin{aligned}
p_{\{1\}} + p_{\{2\}} + p_{\{1,2\}} &\leq p_1 + p_2 \\
&\leq p_{\{1\}} + p_{\{2\}} + p_{\{1,2\}} + p_{\{2,3\}} + p_{\{1,2,3\}} \\
p_{\{2\}} + p_{\{3\}} + p_{\{2,3\}} &\leq p_2 + p_3 \\
&\leq p_{\{2\}} + p_{\{3\}} + p_{\{1,2\}} + p_{\{2,3\}} + p_{\{1,2,3\}}
\end{aligned}$$

and

$$\begin{aligned}
p_{\{1\}} &\leq p_1 \leq p_{\{1\}} + p_{\{1,2\}} + p_{\{1,2,3\}} \\
p_{\{2\}} &\leq p_2 \leq p_{\{2\}} + p_{\{1,2\}} + p_{\{2,3\}} + p_{\{1,2,3\}} \\
p_{\{1\}} + p_{\{2\}} + p_{\{1,2\}} &\leq p_1 + p_2 \\
&\leq p_{\{1\}} + p_{\{2\}} + p_{\{1,2\}} + p_{\{2,3\}} + p_{\{1,2,3\}} \\
p_{\{2\}} + p_{\{2,3\}} &\leq p_2 + p_3 \\
&\leq p_{\{2\}} + p_{\{1,2\}} + p_{\{2,3\}} + p_{\{1,2,3\}}
\end{aligned}$$

It is instructive to use a few numerical values for the distribution of Y^* to get an impression about the dependence of the range of distributions of Y implied by restrictions on the support of Y^*. For the three examples, I will use the following values:

Example 1

$$\mathcal{Y}_1^* = \{\{1\}, \{2\}, \{3\}, \{2,3\}, \{1,2,3\}\}$$
$$p_{\{1\}} = 0.5, p_{\{2\}} = 0.1, p_{\{3\}} = 0.1, p_{\{2,3\}} = 0.1, p_{\{1,2,3\}} = 0.2$$

Example 2

$$\mathcal{Y}_2^* = \{\{1\}, \{2\}, \{3\}, \{1,2\}, \{2,3\}, \{1,2,3\}\}$$
$$p_{\{1\}} = 0.5, p_{\{2\}} = 0.1, p_{\{3\}} = 0.1, p_{\{1,2\}} = 0.1, p_{\{2,3\}} = 0.1,$$
$$p_{\{1,2,3\}} = 0.1$$

Example 3

$$\mathcal{Y}_3^* = \{\{1\}, \{2\}, \{1,2\}, \{2,3\}, \{1,2,3\}\}$$
$$p_{\{1\}} = 0.5, p_{\{2\}} = 0.1, p_{\{1,2\}} = 0.1, p_{\{2,3\}} = 0.1, p_{\{1,2,3\}} = 0.2$$

A closer look at the first example provides further insight into the structure implied by the consistency condition. Each of the inequalities defines a closed half-space in $\mathbb{R}^3$. The set of points satisfying all the inequalities is the intersection of these half-spaces. Thus the set satisfying the consistency condition must be a polytope (Ziegler 1995: Chap. 1).[8] Since all the lower and upper bounds for (p_1, p_2, p_3) are in the interval [0,1], the set of points satisfying the first three constraints is a closed cuboid with sides parallel to the axes in the unit cube. The last constraint may chop off the outer edges parallel to the first axes. The polytope of solutions is depicted in Figure 2.5.

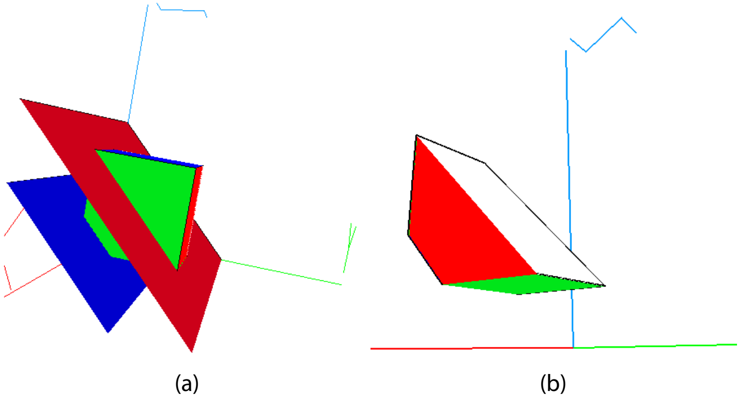


Figure 2.5.: The polytope of the consistency conditions of Y_1^*. The first figure presents the constraints of the last line of the conditions. The second figure shows the resulting polytope. The distribution of Y^* is given by $p_{\{1\}} = 0.5, p_{\{2\}} = 0.1, p_{\{3\}} = 0.1, p_{\{2,3\}} = 0.1, p_{\{1,2,3\}} = 0.2$.

The intersection of this polytope with the simplex $p_1 + p_2 + p_3 = 1, p_i \geq 0$ gives the set of probability distributions compatible with the constraints. The intersection could only be empty if the polytope would lie completely on either side of the simplex. This can be checked by computing the vertices of the polytope from the constraints given in (2.15). Using the probabilities suggested above, the vertices are

[8] Ziegler's book (1995) provides a vivid introduction to the theory of convex polytopes.

(0.5, 0.1, 0.2)	(0.5, 0.1, 0.4)	(0.7, 0.1, 0.2)	(0.7, 0.4, 0.1)
(0.5, 0.2, 0.1)	(0.5, 0.4, 0.1)	(0.7, 0.2, 0.1)	(0.7, 0.1, 0.4)

Here, the first two points on the left lie below the probability simplex, the last two on the right above it. Thus the set of probability distributions subject to the constraints given by the consistency requirements (2.15) is non-empty. As an intersection of half-spaces with the probability simplex, it is a closed polytope. In particular, the solution set is convex. The set of probability distributions of Y compatible with the distribution of Y^* given above is depicted in Figure 2.6. The resulting set of probability distributions p_i is shown in Figure 2.7 together with the solution sets of the other two examples.

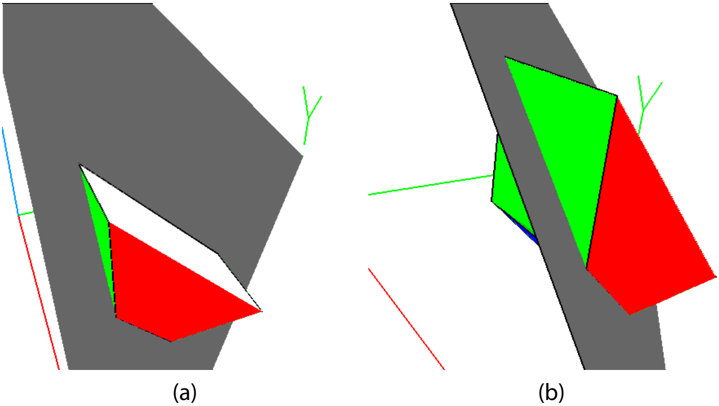


Figure 2.6.: Two views of the constraints implied by the distribution of Y_1^* intersecting the probability simplex.

Geometrically, the inequalities of the consistency requirements generally lead to a closed convex polytope as the set of compatible values of the distribution of Y. Requiring additionally that the values are in fact a probability distribution translates into intersecting the polytope with the simplex of probability distributions giving ones again a closed and convex polytope. The extreme points of this polytope are in Example 1

$$\begin{array}{cc} (0.5, 0.1, 0.4) & (0.5, 0.4, 0.1) \\ (0.7, 0.2, 0.1) & (0.7, 0.1, 0.2) \end{array}$$

In the other two examples, the extreme points of the polytope turn out to be

$$\begin{array}{cc} (0.5, 0.2, 0.3) & (0.6, 0.1, 0.3) \\ (0.5, 0.4, 0.1) & (0.7, 0.2, 0.1) \\ (0.7, 0.1, 0.2) & \end{array} \quad \text{and} \quad \begin{array}{cc} (0.6, 0.1, 0.3) & (0.8, 0.2, 0.0) \\ (0.5, 0.2, 0.3) & (0.5, 0.5, 0.0) \\ (0.8, 0.1, 0.1) & \end{array}$$

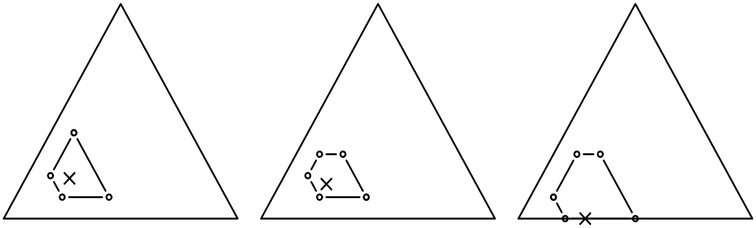


Figure 2.7.: The ranges of (p_1, p_2, p_3) implied by $\mathcal{Y}_1^*, \mathcal{Y}_2^*, \mathcal{Y}_3^*$ and the following probabilities: (0.5, 0.1, 0.1, 0.1, 0.2), (0.5, 0.1, 0.1, 0.1, 0.1, 0.1), and (0.5, 0.1, 0.1, 0.1, 0.2). The coarsening at random solutions are indicated by $\times$.

As an immediate consequence of this geometric interpretation of the consistency conditions, a measure of the informativeness of the distribution of Y^* for the distribution of Y suggests itself: The volume of the solution set in relation to the volume of the simplex. In the three examples above, the volume of the possible distributions relative to the volume of the simplex is 8%, 7%, and 14%, respectively.[9] This may be compared with the simple version of missing versus exact answers in the case of the

[9] The computation of the volumes depends heavily on 'nice' triangulations of the solution sets, and these may be difficult to find. Furthermore, constructing the solution sets becomes rather complicated both with an increasing number of constraints and an increasing number of elements of $\mathcal{Y}$. The general formulation (2.13) suggests that the complexity may increase exponentially with the size of $\mathcal{Y}$. Even in very restricted cases, where the number of binding inequalities may be much less than (2.13), programs such as `polymake` (`www.math.tu-berlin.de/polymake`) must be used

National Longitudinal Survey of Youth where the observed information reduces the possible range of percentages and thus the one dimensional volume to only 18%.

With dichotomous variables, it is only possible to distinguish between either exact observations or no answer at all. In that case, the simple bounds of the introductory example characterise the set consistent distributions. But with larger numbers of states $|\mathcal{Y}|$, the complexity of incomplete data models increases. At the same time, using the partial information (not just assuming that partial, inexact information is completely missing) may dramatically increase the information that can be extracted from the distribution on $\mathcal{Y}^*$. To illustrate the possible gain from using also the partial information, Figure 2.8 compares the ranges of consistent distributions implied by the distributions on $\mathcal{Y}_1^*, \mathcal{Y}_2^*, \mathcal{Y}_3^*$ with those that arise when all partial information is treated as completely missing. In the latter case, the set of consistent distributions is always a simplex whose vertices are the probabilities of the singletons in $\mathcal{Y}^*$ where the mass of the missing data is added to each element in turn. The effect of including the partial information is to reduce the range of compatible distributions by excluding the grey areas from consideration. The gain of information (the relative reduction of the area of consistent distributions) is 12%, 28%, and 14%, respectively.

2.1.5. Constructing Consistent Distributions

The consistency conditions in form of the inequalities (2.12) or (2.13) provide a simple and intuitive view of the form of the implied set of distributions of Y as an intersection of half-spaces with the unit-simplex. Before one can use the result, however, it must be checked whether the solution set always is nonempty. That is, given a structure $\mathcal{Y}^*$ and an arbitrary probability distribution on it, is there always at least one

to compute the solution sets. polymake, in turn, relies on cdd (ftp://ftp.ifor.math.ethz.ch/pub/fukuda/cdd/cddman/cddman.html) which may be used to compute the vertices of a polytope from its description in terms of an intersection of half-spaces as in (2.13). See the cdd home page for details on the algorithms employed.

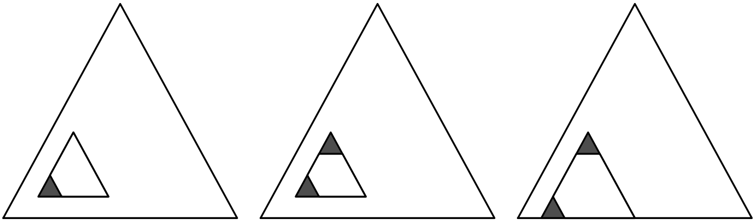


Figure 2.8.: The ranges of (p_1, p_2, p_3) implied by $\mathcal{Y}_1^*, \mathcal{Y}_2^*, \mathcal{Y}_3^*$ compared to the ranges implied by taking partial information as completely missing.

distribution on $\mathcal{Y}$ consistent with the distribution on $\mathcal{Y}^*$? While the geometric interpretation in terms of half-spaces is suggestive, arguments in terms of explicit expressions for the vertices would make the answer obvious. But the connection between the half-space representation of polytopes and their representation as convex hull of their vertices is not immediate. Fortunately, in the case of finite $\mathcal{Y}$, a direct argument can be given: The simplest version would be to fix arbitrary numbers $p(y \mid y^*)$ such that they form, for fixed y^*, a probability distribution on $\mathcal{Y}$ and such that $p(y \mid y^*) = 0$ for $y \notin y^*$. One would then have to check that the implied probabilities $\Pr(Y \in A) \coloneqq \sum_{y \in A} \sum_{y^* \in \mathcal{Y}^*} p(y \mid y^*) \Pr(Y^* = y^*)$ satisfy the consistency constraints (2.11) or (2.13).

While this is rather easy, a somewhat more involved construction gives probability distributions that equal the lower or upper bounds for some set A and otherwise satisfy the constraints. This construction will give further insight into extreme members of the set of probabilities consistent with some distribution of Y^*. For a fixed set $\emptyset \neq A \subseteq \mathcal{Y}$ put

$$\pi_A(y) \coloneqq \begin{cases} \displaystyle\sum_{y^* \subseteq A} \mathbb{1}[y^*](y) \frac{\Pr(Y^* = y^*)}{|y^*|} & \text{for } y \in A \\ \displaystyle\sum_{y^* \nsubseteq A} \mathbb{1}[y^*](y) \frac{\Pr(Y^* = y^*)}{|A^c \cap y^*|} & \text{for } y \notin A \end{cases} \tag{2.16}$$

In Example 1 with $A = \{1\}$, one obtains

$$\pi_{\{1\}}(1) = p_{\{1\}} = 0.5$$

$$\pi_{\{1\}}(2) = p_{\{2\}} + \frac{p_{\{2,3\}}}{2} + \frac{p_{\{1,2,3\}}}{2} = 0.25$$
$$\pi_{\{1\}}(3) = p_{\{3\}} + \frac{p_{\{2,3\}}}{2} + \frac{p_{\{1,2,3\}}}{2} = 0.25$$

and with $A = \{2,3\}$, the definition yields

$$\pi_{\{2,3\}}(1) = p_{\{1\}} + p_{\{1,2,3\}} = 0.7$$
$$\pi_{\{2,3\}}(2) = p_{\{2\}} + \frac{p_{\{2,3\}}}{2} = 0.15$$
$$\pi_{\{2,3\}}(3) = p_{\{3\}} + \frac{p_{\{2,3\}}}{2} = 0.15$$

Thus $\pi_{\{1\}}$ attains the lower bound in the first equation in (2.15) while $\pi_{\{2,3\}}$ attains the upper bound. Also, $\pi_{\{2,3\}}$ attains the lower bound in the last line in (2.15) while $\pi_{\{1\}}$ attains the upper bound. The other inequalities are fulfilled by both $\pi_{\{1\}}$ and $\pi_{\{2,3\}}$. The second and third sets of inequalities are attained by $\pi_{\{2\}}$ and $\pi_{\{3\}}$, and $\pi_{\{1,3\}}$ and $\pi_{\{1,2\}}$, respectively.

In the general case, it must be demonstrated that

a) (2.16) defines a probability density on $\mathcal{Y}$ for all $\emptyset \neq A \subseteq \mathcal{Y}$.

b) $F(A) = \sum_{y \in A} \pi_A(y) =: \Pi_A(A)$, i.e. the bound for A in (2.13) is attained by choosing π_A as a density on $\mathcal{Y}$. Here I write $\Pi_A(.)$ for the measure corresponding to the density $\pi_A(.)$.

c) The density π_A satisfies all other consistency conditions in that for all subsets $\emptyset \neq B \subseteq \mathcal{Y}$

$$F(B) \leq \sum_{y \in B} \pi_A(y) = \Pi_A(B)$$

First note that

$$\Pi_A(\mathcal{Y}) = \Pi_A(A) + \Pi_A(A^c)$$

or in terms of the density

$$\sum_{y \in \mathcal{Y}} \pi_A(y) = \sum_{y \in A} \pi_A(y) + \sum_{y \notin A} \pi_A(y)$$

For the first term on the right hand side

$$\begin{aligned}
\sum_{y\in A} \pi_A(y) &= \sum_{y\in A} \sum_{y^*\subseteq A} \mathbb{1}[y^*](y) \frac{\Pr(Y^* = y^*)}{|y^*|} \\
&= \sum_{y^*\subseteq A} \frac{\Pr(Y^* = y^*)}{|y^*|} \sum_{y\in y^*} 1 \\
&= \sum_{y^*\subseteq A} \Pr(Y^* = y^*) = F(A)
\end{aligned}$$

where in the second equation the order of summation is interchanged. It follows that the lower bound in the consistency condition (2.13) is attained for the set A by this construction. Thus requirement b) holds true. Further,

$$\begin{aligned}
\sum_{y\notin A} \pi_A(y) &= \sum_{y\notin A} \sum_{y^*\nsubseteq A} \mathbb{1}[y^*](y) \frac{\Pr(Y^* = y^*)}{|A^c \cap y^*)|} \\
&= \sum_{y} \sum_{y^*\nsubseteq A} \mathbb{1}[A^c](y)\, \mathbb{1}[y^*](y) \frac{\Pr(Y^* = y^*)}{|A^c \cap y^*)|} \\
&= \sum_{y^*\nsubseteq A} \sum_{y} \mathbb{1}[A^c \cap y^*](y) \frac{\Pr(Y^* = y^*)}{|A^c \cap y^*|} \\
&= \sum_{y^*\nsubseteq A} \Pr(Y^* = y^*)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\Pi_A(\mathcal{Y}) &= \sum_{y\in\mathcal{Y}} \pi_A(y) = \sum_{y\in A} \pi_A(y) + \sum_{y\notin A} \pi_A(y) \\
&= \sum_{y^*\subseteq A} \Pr(Y^* = y^*) + \sum_{y^*\nsubseteq A} \Pr(Y^* = y^*) = 1
\end{aligned}$$

and, since certainly $\pi_A(y) \geq 0$, $\pi_A(.)$ is a probability density on $\mathcal{Y}$ (requirement a). It remains to show that π_A satisfies all the other constraints implied by the consistency conditions (2.13). Let $\emptyset \neq B \subset \mathcal{Y}$ be another set that might appear in the consistency conditions. Writing $\Pi_A(B) \coloneqq \sum_{y \in B} \pi_A(y)$ for the measure corresponding to π_A,

$$\Pi_A(B) = \Pi_A(A \cap B) + \Pi_A(A^c \cap B)$$

Now

$$\begin{aligned}
\Pi_A(A \cap B) &= \sum_{y \in A \cap B} \sum_{y^* \subseteq A} \mathbb{1}[y^*](y) \frac{\Pr(Y^* = y^*)}{|y^*|} \\
&= \sum_{y^* \subseteq A} \sum_{y \in A \cap B} \mathbb{1}[y^*](y) \frac{\Pr(Y^* = y^*)}{|y^*|} \\
&\geq \sum_{y^* \subseteq A \cap B} \sum_{y \in A \cap B} \mathbb{1}[y^*](y) \frac{\Pr(Y^* = y^*)}{|y^*|} \\
&= \sum_{y^* \subseteq A \cap B} \Pr(Y^* = y^*)
\end{aligned}$$

and

$$\begin{aligned}
\Pi_A(A^c \cap B) &= \sum_{y \in A^c \cap B} \sum_{y^* \not\subseteq A} \mathbb{1}[y^*](y) \frac{\Pr(Y^* = y^*)}{|A^c \cap y^*|} \\
&= \sum_{y^* \not\subseteq A} \sum_{y \in A^c \cap B} \mathbb{1}[y^*](y) \frac{\Pr(Y^* = y^*)}{|A^c \cap y^*|} \\
&\geq \sum_{\substack{y^* \not\subseteq A \\ y^* \subseteq B}} \sum_{y \in A^c \cap B} \mathbb{1}[y^*](y) \frac{\Pr(Y^* = y^*)}{|A^c \cap y^*|} \\
&= \sum_{\substack{y^* \not\subseteq A \\ y^* \subseteq B}} \Pr(Y^* = y^*)
\end{aligned}$$

so that

$$\begin{aligned}
\Pi_A(B) &= \Pi_A(A \cap B) + \Pi_A(A^c \cap B) \\
&\geq \sum_{y^* \subseteq A \cap B} \Pr(Y^* = y^*) + \sum_{\substack{y^* \not\subseteq A \\ y^* \subseteq B}} \Pr(Y^* = y^*)
\end{aligned}$$

$$= \sum_{y^* \subseteq B} \Pr(Y^* = y^*) = F(B)$$

Hence, also requirement c) is satisfied.

It follows that to every structure of incomplete data $\mathcal{Y}^*$ and to every distribution of the corresponding Y^* there exists a non-empty set of probability distributions on $\mathcal{Y}$. The set is restricted only by the distributional consequences of the consistency requirement $Y(.) \in Y^*(.)$. Geometrically, the set of consistent distributions of Y is a non-empty, closed, convex polytope.

This answers one of the questions on the "simple device" formulated in the introduction, namely whether by introducing a probabilistic framework the set of underlying possible values consistent with the incomplete data is restricted beyond the consistency requirement. This does not happen, the "simple device" does not prejudice certain solutions against others simply by adding the constraints of a probability model.

2.1.6. Selectors and Allocations

But the previous construction gives only a marginal distribution consistent with the distributional consequences of the requirement $Y \in Y^*$. What must be shown in order to be sure that there is no systematic bias introduced by adopting a probability model is that the requirement $Y(\omega) \in Y^*(\omega)$ for almost all $\omega \in \Omega$ is at most as strong as the distributional consistency requirement (2.13). Note that when dealing with statistical variables as functions of population members, the question is one pertaining to facts. It may be simply wrong that $Y(u) \in Y^*(u)$ for all $u \in \mathcal{U}$. But given $Y(u) \in Y^*(u)$ as a matter of fact, all (empirical) marginal distributions of Y satisfying (2.13) are possible distributions of interest with no further restrictions.

However, in a probability model one must be able to construct random variables (Y, Y^*) defined on a common probability space $(\Omega, \mathcal{B}, \lambda)$ such that $\Pr(Y \in Y^*) = 1$ and such that the marginal distribution is prescribed by $\Pr(Y^* = y^*)$ for all $y^* \in \mathcal{Y}^*$. If such a construction succeeds, the constructed random variable Y is called a *selector* of the set-valued variable Y^*. Thus Y is a selector of Y^* if and only if Y^* is a coarsening of Y.

The construction in the previous subsection exhibits one possible distribution on $\mathcal{Y}$ consistent with the distributional consequences of the requirement $Y(.) \in Y^*(.)$. Moreover, $\Pi_A(A)$ attains the lower bound $F(A)$. But comparing the values of π_A in the example with the vertices of the polytope of consistent distributions shows that these are not the extreme points of the polytope. Thus, a concise description of the polytope of consistent distributions is also still missing.

Both problems, that of the extent of the set of probability models satisfying $\Pr(Y \in Y^*) = 1$ and that of characterising the set of distributionally feasible marginal distributions, can be attacked using a very simple device. Let

$$\alpha \colon \mathcal{Y} \times \mathcal{P}(\mathcal{Y}) \longrightarrow [0,1] \tag{2.17}$$

such that

$$\Pr(Y^* = y^*) = \sum_{y \in y^*} \alpha(y, y^*) \qquad \forall\, y^* \in \mathcal{Y}^* \tag{2.18}$$

Such a function is called an *allocation* in the literature on random sets. It is an obvious candidate for a joint density of (Y, Y^*), at least when $\alpha(y, y^*) = 0$ for $y \notin y^*$.[10] In fact, if one defines the marginal density of Y as

$$\pi_\alpha(y) \coloneqq \sum_{y^* \ni y} \alpha(y, y^*) = \sum_{y^* \in \mathcal{Y}^*} \mathbb{1}[y^*](y)\alpha(y, y^*)$$

the marginal density of Y^* stays just $\Pr(Y^* = y^*)$. Furthermore,

$$\sum_{y^* \in \mathcal{Y}^*} \sum_{y \in \mathcal{Y}} \mathbb{1}[y^*](y)\alpha(y, y^*) = \sum_{y^* \in \mathcal{Y}^*} \sum_{y \in y^*} \alpha(y, y^*)$$

[10] It is sometimes convenient not to require $\alpha(y, y^*) = 0$ for $y \notin y^*$. In that case, α need not be a density. It can be made into one by multiplying α by $\mathbb{1}[y^*](y)$ since this modification does not change the defining property (2.18). This is the reason to call α an allocation, not a joined density. Whether or not $\alpha(y, y^*) = 0$ for $y \notin y^*$ is assumed should be clear from the context.

$$= \sum_{y^* \in \mathcal{Y}^*} \Pr(Y^* = y^*) = 1$$

Thus $\mathbb{1}[y^*](y)\alpha(.,.)$ is a joined density of (Y, Y^*) with the prescribed marginal distribution of Y^*. Moreover,

$$\Pr(Y \notin Y^*) = \sum_{y* \in \mathcal{Y}^*} \sum_{y \notin y^*} \mathbb{1}[y^*](y)\alpha(y, y^*) = 0$$

so that $\Pr(Y \in Y^*) = 1$. In particular, the marginal measure Π_α derived from the density π_α satisfies the consistency equations (2.13) so that $F(A) \leq \Pi_\alpha(A)$ for all $A \subseteq \mathcal{Y}$.

By the above construction, there always exists a common probability space $(\Omega, \mathcal{B}, \lambda)$ for two variables (Y, Y^*) such that Y is a selector for Y^* and Y^* has any given marginal distribution. In particular, one may choose $\Omega = \mathcal{Y} \times \mathcal{P}(\mathcal{Y})$, $\mathcal{B} = \mathcal{P}(\Omega)$ and the density of λ as $\mathbb{1}[.](.)\alpha(.,.)$ for any allocation $\alpha(.,.)$. The pair of random variables (Y, Y^*) can then be taken as the identity $(Y, Y^*)(\omega) = (Y, Y^*)((y, y^*)) \coloneqq (y, y^*)$. In probability theory, such constructions of pairs of random variables on a common probability space are called *couplings*.[11]

It remains to exhibit allocations and, if possible, construct some versions with properties that elucidate the structure of incomplete data problems. A simple construction starts out with any probability density, say $\pi(.)$ on $\mathcal{Y}$ such that $\pi(y) > 0$ for all y. Then one may put

$$\alpha(y, y^*) \coloneqq \pi(y) \frac{\Pr(Y^* = y^*)}{\Pi(y^*)} \tag{2.19}$$

where Π is the measure corresponding to the density π. Note that with this definition $\alpha(y, y^*) > 0$ for all y as long as $\Pr(Y^* = y^*) > 0$. To provide a joint density, one may simply put

$$\alpha'(y, y^*) \coloneqq \mathbb{1}[y^*](y)\alpha(y, y^*)$$

[11] See Thorisson (2000) or Lindval (1992) who both provide lucid examples and a wealth of further constructions.

This construction will be of particular importance when different versions of the CAR condition are discussed. To illustrate it with the data from the first example, set $\pi(y) \coloneqq 1/3$ for all $y \in \mathcal{Y} = \{1, 2, 3\}$. Then α becomes

	$\{1\}$	$\{2\}$	$\{3\}$	$\{2,3\}$	$\{1,2,3\}$
1	0.5	0.1	0.1	0.05	0.2/3
2	0.5	0.1	0.1	0.05	0.2/3
3	0.5	0.1	0.1	0.05	0.2/3

which clearly is no density. But the corresponding α' is given by

	$\{1\}$	$\{2\}$	$\{3\}$	$\{2,3\}$	$\{1,2,3\}$
1	0.5	0	0	0	0.2/3
2	0	0.1	0	0.05	0.2/3
3	0	0	0.1	0.05	0.2/3

which is a density with the marginal $\pi_{\alpha'}(1) = 0.5 + 0.2/3$ and $\pi_{\alpha'}(2) = \pi_{\alpha'}(3) = 0.1 + 0.2/3$ which clearly satisfies (2.15).

Thus, there always exists at least one pair of random variables (Y, Y^*) that the "simple device" requires for its operation. For finite $\mathcal{Y}$, the existence of such a coupling satisfying the additional consistency requirement is almost trivial. It is, of course, always possible to choose an element from a finite, non-empty set as is required here.[12]

But it is not yet clear how large the set of possible joined distributions is. In the case of statistical variables, it coincides with the set of consistent distributions characterised in the previous sections since $Y(u) \in Y^*(u)$ as a matter of fact. Either it is the case that all respondents answer

[12] The proof of the existence of almost sure selectors for infinite (even countable) $\mathcal{Y}$ is much more involved. A direct construction of a selector will not work. Nevertheless, for general Polish spaces, the existence of selectors is guaranteed by the Kuratowski-Ryll-Nardzewski selection theorem. Moreover, in this setting, one has to be prepared to deal with the dependence of the set of allocations on the particularly chosen underlying probability space. Molchanov (2005: 32) provides an example. In consequence, the "simple device" has to deal with classes of selectors that are not unique. It turns out, however, that the (weak) closure of the set of selectors is in fact uniquely defined.

consistently, or it is not. But consistency is not automatically guaranteed for random variables. If the "simple device" is to work, the use of random variables (Y, Y^*) should not restrict the set of compatible marginal distributions of Y for a given marginal distribution of Y^* beyond the distributional requirements (2.13). Put differently, one has to show that the marginal distributions of all possible selectors of a random set Y^* with a prescribed distribution exhausts the set of consistent distributions.

One possible approach is to investigate the set of all allocations, preferably constructing extreme ones that in some (vague) sense chart the boundary of the set of all allocations. Instead of distributing the mass $\Pr(Y^* = y^*)$ according to some probability π on the elements of $\mathcal{Y}$ one may try to distribute as much of the mass to a selected point, say $y_{|\mathcal{Y}|}$, distribute as much as possible of the remaining mass on a second point, say $y_{|\mathcal{Y}|-1}$, and so on. More precisely, suppose that $\mathcal{Y}$ is ordered as $(y_1, y_2, \ldots, y_{|\mathcal{Y}|})$. Then put $\alpha(y, y^*) := \Pr(Y^* = y^*)$ if y is the maximal element in y^* according to this ordering and set $\alpha(y, y^*) = 0$ for all other y. Then the marginal density of Y becomes

$$\begin{aligned}\pi_\alpha(y_i) &= \sum_{y^* \ni y_i} \alpha(y_i, y^*) = \sum_{\substack{y^* \ni y_i \\ y^* \subseteq \{y_1, y_2, \ldots, y_i\}}} \Pr(Y^* = y^*) \\ &= F(\{y_1, y_2, \ldots, y_i\}) - F(\{y_1, y_2, \ldots, y_{i-1}\})\end{aligned} \tag{2.20}$$

where as before $F(.)$ is the distribution function of Y^* with $F(A) = \Pr(Y^* \subseteq A)$.

With the ordering $(1, 2, 3)$, this construction in Example 1 leads to the joint density

	$\{1\}$	$\{2\}$	$\{3\}$	$\{2,3\}$	$\{1,2,3\}$
1	0.5	0	0	0	0
2	0	0.1	0	0	0
3	0	0	0.1	0.1	0.2

so that the marginal density becomes $\pi_\alpha(1) = 0.5$, $\pi_\alpha(2) = 0.1$, $\pi_\alpha(3) = 0.4$. Comparing this with the extreme points of the feasible distributions given in section 2.1.4, this solution is extreme in the set of consistent distributions. In fact, computing π_α for each permutation of $(1, 2, 3)$ leads to all the extreme points of the polytope of consistent distributions: The ordering $(2, 1, 3)$ leads to the same π_α as that arising from $(1, 2, 3)$, the orderings $(1, 3, 2)$ and $(3, 1, 2)$ both lead to $\pi_\alpha = (0.5, 0.4, 0.1)$, and the orderings $(2, 3, 1)$ and $(3, 2, 1)$ lead to $\pi_\alpha = (0.7, 0.1, 0.2)$ and $(0.7, 0.2, 0.1)$

Now the construction (2.20) depends on the ordering of the elements of $\mathcal{Y}$. For each permutation σ of $(y_1, y_2, \ldots, y_{|\mathcal{Y}|})$ one gets an allocation and a corresponding marginal density of Y which one might index with the permutation σ, say π_σ. It can be shown that all π_σ are extreme points of the set of densities π on $\mathcal{Y}$ that satisfy the consistency requirements $F(A) \leq \Pi_\sigma(A)$ for all $A \subseteq \mathcal{Y}$. Moreover, there are only these extreme points.[13]

This result has many important consequences. First note that one gets an upper bound on the number of extreme points as $|\mathcal{Y}|!$. But of much greater importance is the following consequence: The set $\{\Pi \mid F(A) \leq \Pi(A) \forall A \subseteq \mathcal{Y}\}$ of measures that satisfy the (distributional) consistency conditions is equal to the set of distributions that come from some allocation $\alpha(.,.)$. This is because by construction, $\{\Pi_\alpha \mid \alpha \text{ is an allocation}\} \subseteq \{\Pi \mid F(A) \leq \Pi(A) \forall A \subseteq \mathcal{Y}\}$. On the other hand, since $\{\Pi_\sigma \mid \sigma\}$ are the extreme points of $\{\Pi \mid F(A) \leq \Pi(A) \forall A \subseteq \mathcal{Y}\}$, the latter set is the convex hull $\operatorname{conv}(\{\Pi_\sigma\})$ of the extreme allocations. But convex combinations of allocations are again allocations so that also $\{\Pi \mid F(A) \leq \Pi(A) \forall A \subseteq \mathcal{Y}\} \subseteq \{\Pi_\alpha \mid \alpha \text{ is an allocation}\}$. Therefore to each of the measures Π satisfying (2.13) there exists a pair of random variables (Y, Y^*) where Y is a selector for Y^* and such that the marginal measure of Y is Π.

This answers another possible concern with the “simple device”: Using

[13] See Nguyen (2006: 98–102, 118–122), or Nguyen/Wu (2006) and Aubin and Frankowska (1990: Chap. 8). Feng/Feng (2004) generalise the result to compact metric spaces with continuous distributions. Such generalisations are also discussed in Molchanov (2005: Chap. 1).
However, this construction of the vertices (extreme points) of the feasible set of marginal distributions on $\mathcal{Y}$ is mainly of theoretical importance. It is numerically much faster and requires less memory to use the classical double description method from the theory of linear optimisation.

statistical variables, the set of possible distributions of Y consistent with the observations Y^* is just the set of distributions satisfying $F(A) \leq \Pr(Y \in A)$ since $Y(u) \in Y^*(u)$ for all u as a matter of fact. But it is now clear that the same applies to a probability model. Thus, the set of probability models $(\Omega, \mathcal{B}, \lambda, (Y, Y^*))$ compatible with some prescribed distribution of Y^* is only constrained by the same requirement that must hold for statistical variables as well and by nothing beyond.

2.2. Coarsening at Random

Problems of incomplete data and particularly the special case of either completely missing or exactly observed data have been discussed since the inception of statistics. But such problems were dealt with using either special extensions to algorithms developed within the framework of classical statistics (and typically restricted to completely missing observations) or by exploiting the structure of special types of incomplete data. The classical approach in the former tradition is well summarised in a series of articles by Afifi and Elashoff (1966, 1967, 1969). The treatment of partially complete observations can be traced back to the very first contributions to statistics in its present form. E.g., Ronald Fisher in his famous essay on maximum likelihood of 1922 deals at length with the problem of grouped data. As far as I know, however, a systematic treatment unifying the different approaches was never attempted before the 1990's.

The introduction of the "simple device" had to wait until 1976 when Rubin's account appeared in Biometrika. Xiao-Li Meng's editorial for the Statistica Sinica special issue on missing data (vol. 16, no. 4, 2006) recounts details of the reservations of referees and editors before the paper was finally accepted. The resistance the paper met early on was certainly not only based on technicalities, as Meng seems to imply. As I indicated in the introduction, scepticism towards the "simple device" was widespread and well articulated.

Rubin not only gave a systematic account of missing data based on the "simple device", he also introduced the notion of 'missing at random' (MAR), a condition on the joint distribution of (Y, Y^*) that allows statistical methods to proceed as if no observations were missing. He also discussed the consequences of the MAR condition for inference within the different statistical paradigms. The first contribution made the "simple device" very attractive to applied statisticians, for obvious reasons. The general approach also sparked the interest of most statisticians and probabilists since it promised a principled treatment where previously ad hoc methods and subjective judgements prevailed. Thus, Rubin's 1976 paper marks the starting point of the modern treatment of missing data problems.

Even though the "simple device" was accepted rather rapidly and even though some of the techniques developed within the framework became common practice in many applied fields, it took another 15 years until the approach was extended to partially complete observations. A first attempt at a general formulation was made by Heitjan and Rubin in 1991. They termed conditions similar to the MAR conditions *coarsening at random* (CAR) conditions. The notion was further developed by Heitjan (1993, 1994), Jacobsen and Keiding (1995), Gill et al. (1997), Nielsen (2000), Grünwald and Halpern (2003), Jaeger (2005a, b), Lu and Copas (2004), de Cooman and Zaffalon (2004), and by Cator (2004).[14]

In the last section I have shown that a versatile representation of partially incomplete data can be obtained when incomplete data are taken to be subsets of the range of statistical or random variables. If a variable Y takes values in $\mathcal{Y}$, then partial knowledge of the value of the variable can be encoded by the subset of $\mathcal{Y}$ to which it necessarily belongs. Thus, incomplete data can be represented as derived variables Y^* with values in the power set $\mathcal{P}(\mathcal{Y})$, excluding the empty set. Such a representation presupposes that for all elements either of a population $\mathcal{U}$ or a probability field Ω, $Y(.) \in Y^*(.)$. In the case of data for a population, this consistency condition is satisfied as a matter of fact. In the case of a probability model, it was demonstrated in the previous section that the consistency requirement can always be satisfied for any given distribution of the variable Y^*. Thus one can always construct an abstract probability space on which both Y and Y^* are defined. In particular, one may always assume that there is a joint distribution of Y and Y^*.

The missing at random condition of Rubin as well as the general coarsening at random condition is formulated as an additional constraint

[14] While this rather late development may seem surprising, it may have been due to the fact that partially complete data have been discussed under different headings in different applications, every discipline developing their own jargon. In fact, even today the analysis of censored, grouped, truncated or weighted data is often presented in monographs and articles accessible only to experts in a particular subject area. But this explanation is partial at best, since many unifying concepts were taken up rapidly in other areas of statistics.

on this joint distribution. It can be expressed most conveniently using conditional probabilities. One of its versions requires that for all y^* and all $y \in y^*$:

$$\Pr(Y = y \mid Y^* = y^*) = \Pr(Y = y \mid Y \in y^*) \tag{2.21}$$

In words, the answers of respondents, y^*, just tell that the random variable Y takes a value in y^* and nothing more. Or in Bayesian terms, the fact that $\{Y^* = y^*\}$ was observed allows to update probabilities only to the same extent as an observation of $\{Y \in y^*\}$ and does not provide any further clues.

The marginal distribution of Y can be obtained from combining the conditional distribution of Y given Y^* with the marginal distribution of Y^*, i.e.

$$\Pr(Y = y) = \sum_{y^* \in \mathcal{Y}^*} \Pr(Y = y \mid Y^* = y^*) \Pr(Y^* = y^*)$$

but from (2.21), this equals

$$\begin{aligned} \Pr(Y = y) &= \sum_{y^* \in \mathcal{Y}^*} \Pr(Y^* = y^*) \Pr(Y = y \mid Y \in y^*) \\ &= \sum_{\substack{y^* \in \mathcal{Y}^* \\ y^* \ni y \\ \Pr(Y^* = y^*) > 0}} \Pr(Y^* = y^*) \frac{\Pr(Y = y)}{\Pr(Y \in y^*)} \end{aligned} \tag{2.22}$$

since terms involving either sets y^* with $y \notin y^*$ or with $\Pr(Y^* = y^*) = 0$ are zero by definition. Furthermore, if $\Pr(Y^* = y^*) > 0$, then because of (2.3), $\Pr(Y \in y^*) > 0$ so that the last term is well defined.

Note that if $\Pr(Y^* = y^*) = 0$ but $\Pr(Y \in y^*) > 0$, then (2.21) may be seen as an arbitrary definition of the left hand side. If both sets $\{Y^* = y^*\}$ and $\{Y \in y^*\}$ have probability 0, then nothing is required.[15] That is,

[15] In particular, the explicit restriction of the CAR requirement (2.21) to sets $\{Y^* = y^*\}$ of positive probability as in Jaeger (2005a: 1969) is superfluous.
Also, in (2.22) one can not, in general, divide out $\Pr(Y = y)$ since it may be zero.

the CAR condition requires that the probability mass $\Pr(Y^* = y^*)$ of a y^* is distributed among the $\{Y = y\}$ according to their conditional distribution given $Y \in y^*$.

Returning to the introductory example of the National Longitudinal Survey of Youth, suppose that an answer to the employment question is either correct or is completely missing. In my notation, missing answers are indicated by the set $y^* = \{0, 1\}$. Suppose that the CAR condition (2.21) holds. Then $\Pr(Y = y \mid Y^* = \{0, 1\}) = \Pr(Y = y \mid Y \in \{0, 1\})$. But since it was supposed that all respondents are either employed or unemployed, the last quantity reduces to $\Pr(Y = y)$. Consequently, to a Bayesian the answer $Y^* = \{0, 1\}$ would not change his prior probabilities $\Pr(Y = y)$. And within all other approaches to statistics, the information $Y^* = \{0, 1\}$ is of no use in "inferential" procedures either. Thus, at least for the case of either completely missing or exact observations the condition (2.21) reproduces Rubin's MAR condition in that completely missing observations can safely be ignored.

In this simple example, the conditional probabilities are

$$\begin{aligned}
\Pr(Y = 1 \mid Y \in \{1\}) &= 1 \\
\Pr(Y = 1 \mid Y \in \{0\}) &= 0 \\
\Pr(Y = 1 \mid Y \in \{0, 1\}) &= \Pr(Y = 1)
\end{aligned}$$

Thus

$$\Pr(Y = 1) = \Pr(Y^* = \{1\}) + \Pr(Y = 1)\Pr(Y^* = \{0, 1\})$$

leading to

$$\Pr(Y = 1) = \frac{\Pr(Y^* = \{1\})}{1 - \Pr(Y^* = \{0, 1\})} = \frac{\Pr(Y^* = \{1\})}{\Pr(Y^* \neq \{0, 1\})} \tag{2.23}$$

Similarly $\Pr(Y = 0) = \Pr(Y^* = \{0\}) / \Pr(Y^* \neq \{0, 1\})$. Within the model the probability of being employed can be deduced form the probability of reporting to be employed.

2.2.1. Does CAR Apply to Survey Non-response?

The textbook approach connecting the probability model with the sampled responses adds the assumption that the $(Y(u,.), Y^*(u,.))$ are identically and independently distributed across the population. It further adds that in some sense the mean value of observations should be close to the expectation derived from a reasonable probability model, i.e.

$$m(Y,S) \coloneqq \frac{1}{|S|}\sum_{u\in S} Y(u) \approx \mathbb{E}(Y(u_1,.)) = \Pr(Y(u_1,.)=1) \qquad (2.24)$$

The precise nature of the connection between random variables and observations is of no immediate concern here. A different or more elaborate approach would lead to essentially the same procedure. The main point is that within the statistical approach, the search for reasonable values for $m(Y,S)$ can be replaced by a search for a reasonable value of the expectation $\mathbb{E}(Y(u_1,.))$ or, equivalently, for the probability $\Pr(Y(u_1,.) = 1)$. One may proceed similarly to identify (approximately) the observed frequencies of the variables Y^* with their probability counterpart.

Since the variables $(Y(u,.), Y^*(u,.))$ are assumed to be independent and identically distributed, it is possible to drastically reduce the notational burden. The reference to the individuals u or the population $\mathcal{U}$ can be dropped completely. And, as is the custom in most applied statistical work, reference to the elements of the probability space ω will also be dropped. The “simple device” leaves us with a very simple mathematical structure, namely the random tupel (Y, Y^*) and its joint distribution. The original problem, that of connecting the employment rate $m(Y,S)$ to the incomplete answers of the respondents, is now replaced by the problem of characterising the extent to which the distribution of the variable Y can be identified from knowledge of the distribution of Y^* alone.

The “simple device” of introducing a probability model plus the missing at random condition can now be translated back into an “answer” to the original question using a translation principle similar to (2.24). This leads to the naive use of only the complete answers from the survey:

$$\begin{aligned}\Pr(Y=1) &= \frac{\Pr(Y^* = \{1\})}{\Pr(Y^* \neq \{0,1\})} \\ &\approx \frac{\frac{1}{|S|}\sum_{u\in S} \mathbb{1}[Y^* = \{1\}](u)}{\frac{1}{|S|}\sum_{u\in S} \mathbb{1}[Y^* \neq \{0,1\}](u)} \\ &= \frac{1}{|S^*|}\sum_{u\in S^*} \mathbb{1}[Y^* = \{1\}](u)\end{aligned}$$

with $S^* \coloneqq \{u \,|\, u \in S \wedge Y^*(u) \neq \{0,1\}\}$. Here, S^* is a subset of the sample with valid responses to the employment question. Using the data from the National Longitudinal Survey of Youth, one in fact arrives at a percentage of 78% employed.

But what is the precise connection between the results of a survey and its translation into a probability model? And what justifies or at least motivates (2.24)? And what status has the CAR ‘assumption’?

The various textbook approaches are rather unclear and even potentially misleading. Even the very first step, the introduction of a probability model that somehow refers to the information provided by the interviewees, is generally glossed over. And if there are detailed accounts at all, the expositions given are neither enlightening nor cogent.

Consider Haavelmo's influential paper “The Probability Approach to Econometrics” where he wrote

> The question is not whether probabilities *exist* or not, but whether—if we proceed *as if* they existed—we are able to make statements about real phenomena that are “correct for practical purposes”. (Haavelmo 1944: 43)[16]

The literature on stochastic models is similarly obscure. E.g. Nelson in his introduction to “Stochastic Modeling” writes:

[16] See Morgan (1990: Chap. 8) for an emphatic account of the papers impact on econometrics. A more critical account is given by Tryfos (2004: Chap. 7).

> When the data are obtained by observing a real system, then we can treat the data *as if* they were simulation output. There is no need to differentiate "real" data from the output of a simulation. …Probability …is useful for deriving statements about the data that a completely specified model *might generate if it were simulated.* (Nelson 2002: 24)

And Bartholomew in his equally influential "Stochastic Models for Social Processes" writes:

> Our contention is not that the employee actually uses a chance device to make the decision but that the group behaves *as if* its individual members did use such a method. The function of probability theory is thus simply to describe observed variability; it carries no implications about the freedom, or otherwise, of human choice. It is a fact of experience that 'choice may mimic chance'. (Bartholomew 1967: 6)

The 'as if' rhetoric dominated also much of the history of probabilistic reasoning in the social sciences.[17] Historically, the domination of the 'as if' rhetoric may not be too surprising in view of the roots of social statistics in the 19th century, where it was assumed by many that nearly all social phenomena might be amenable to probabilistic analysis.[18]

The philosophical literature on models, while quite extensive, is not very useful either.[19] Even those works directly dealing with statistics like

[17] See Rohwer/Pötter 2002b: Part 2.

[18] Hacking's (1988) account of the origin of experimental design and randomisation from curiosity in telepathy provides an interesting case study of the sort of uses made of probability models in the late 19th century.

[19] Frigg and Hartmann (2005) provide a short review. A somewhat more relevant view is expressed by Knuuttila and Voutilainen (2005).
The classical study of the 'as if' rhetoric is Vaihinger's "Philosophie des Als-Ob" (1911). He clearly distinguishes between 'hypotheses' and 'fiction', writing "Der Verifizierung der Hypothese entspricht die Justifizierung der Fiktion. Muß jene durch Erfahrung bestätigt werden, so muß diese gerechtfertigt werden durch die Dienste, welche sie der Erfahrungswissenschaft schließlich leisten." (1911/1923: 91). But he indiscriminately treats all forms of 'fiction' alike, whether it be mathematical or juridical or philosophical concepts. He therefore is unable to explain the role of

Lenhard (2006) take the probability model for granted. In the statistics literature, when models are discussed at all, then the role of probability is not questioned (e.g. Cox 1990 or Lehmann 1990).

On the other hand, survey statisticians were always aware of the problematic connection between the real answers of interviewees and probabilistic models of them. Tore Dalenius' remark cited in the introduction voices that concern.

Many probabilists were also rather dismayed by the obscure 'as if' rhetoric. Even Kolmogorov in a review chapter that was first published in 1956 stated that

> ...there exists no event which is absolutely random; an event is random or is predetermined depending on the connection in which it is considered, but under specific conditions an event may be random in a completely non-subjective sense, i.e., independently of the state of knowledge of any observer. (Kolmogorov 1999: 249)

While Kolmogorov denounced the observer dependence of randomness but stressed that at least the range of "events" considered has to be taken into account, Jaynes insists on the dependends of the notion on "human information":

> Belief in the existence of 'stochastic processes' in the real world; i.e. that the property of being 'stochastic' rather than 'deterministic' is a real physical property of a process, that exists independently of human information, is another example of the mind projection fallacy: attributing one's own ignorance to Nature instead. (Jaynes 2003: 506)

Matheron, whose "Random Sets and Integral Geometry" provided much inspiration for the treatment of incomplete data, also rejects the 'as if' rhetoric:

> When we deal with a unique phenomenon and a probabilistic model, that is a space (Ω, α, P) which is put in

the 'as if' rhetoric within a model.

> correspondences with this unique reality, the same kind of illusion incites us to say that everything happens, after all, as if the realised event had been “drawn at random” according to law *P* in the sample space Ω. But this is a misleadingly clear statement, and the underlying considerations supporting it are particularly inadequate. What is the mechanism of this “random choice” that we invoke ...? This “random draw” myth, for it is one, (in the pejorative sense), is both useless and gratuitous. (Matheron 1989: 23)

He rather emphasises the role of models and their uses and states:

> In fact there is not, nor can there be, any such thing as probability in itself. *There are only probabilistic models.* In other words, randomness is in no way a uniquely defined, or even definable property of the phenomenon itself. It is only a characteristic of the model or models we choose to describe it, interpret it, and solve this or that problem we have raised about it. (Matheron 1989: 4)

Similarly, Dawid (2004: 44) argued:

> I regard “probability” as a *purely theoretical* term, inhabiting the intellectual universe and without any direct physical counterpart... [W]e should regard probabilities as entering our scientific theories as *instrumental* terms, the link between theoretical probabilities and the physical universe being indirect. This approach to interpreting probabilistic models avoids many potential philosophical pitfalls. In particular, by treating probabilities as purely theoretical terms with only indirect implications for the behavior of observables, it is able to eschew deep but ultimately irrelevant and distracting philosophical inquiry into the “true nature of Probability”.

This emphasis on models allows to see the role of probability in a different light, transcending the objectivist-subjectivist divide. Matheron distinguishes several steps in the choice of a model:

> [T]here is first an *epistemological choice*: it has been decided to use probabilistic techniques to represent the phenomenon.... This is a decision, not a hypothesis. It is a *constitutive decision*. (It 'constitutes' the forest as an object of study, it defines the general framework within which we shall operate and determines the choice of the tools we use). ...At this level, we shall speak of a *constitutive model*.... (Matheron 1989: 52)

Therefore, statements made within a probabilistic framework do not directly refer either to social facts or to states of mind. They only work within a probability model. And the latter is not forced upon us by some obscure 'as if' similarity but created by a decision to use such a model. In fact, the classical theory of sampling does not rely on a probability model for the variables (Y, Y^*). Rather, these variables are treated as fixed though unknown quantities ascribable to the individuals in the population. And probability only appears in the picture in as far as it is introduced by the researcher himself, by deliberately using random sampling designs.

There are some further elements of probability models used in the social sciences that can be regarded as constitutive in the sense of Matheron. In particular, in the context of surveys this is the assumption of independence and of identical distribution. They determine the choice of statistical and probabilistic tools but are not determined from any aspect of social facts.[20] As Kolmogorov (1950: 9) noted: "...one of the most important problems in the philosophy of the natural sciences is ...to make precise the premises which would make it possible to regard any given real events as independent." It is argued here that it is more fruitful not to search for such premises but to take them as part of the decision to use a particular type of model. In fact, classical statistical tests of the independence assumption are self-refuting when carried out on a given data set: If a test does not reject independence, then there must be some

[20] Recently, Humphreys (2008) argued similarly that the basic probability setup is a purely mathematical artefact, a mathematical template. He also counts probability models and classes of distributions as such templates. It seems to me, however, that he overlooked the fundamental role of independence assumptions.

dependence in the joint distribution of the data given the observed value of the test statistic.[21]

From this perspective one needs to reconsider the approximation suggested in (2.24). In so far as the left hand side is treated as a function whose behaviour is determined within the model, it is plain that it can only be correct when several caveats are added. In particular, if $\mathcal{Y}$ is finite, then (2.24) is true, but only with high *probability*. That is, (2.24) can only be expressed when probability is presupposed in the expression of the approximation.

If $\mathcal{Y}$ is unbounded, then (2.24) is not even true in probability since for any sample size $|S|$, there exist distributions such that the difference between the two sides can be made as large as one pleases. This is true even if the existence of all moments or similarly strong regularity conditions are stipulated. Simply put $Y = 0$ with probability ϵ, but $Y = (1/\epsilon)^{k+1}$. Then $\mathbb{E}(Y) = (1/\epsilon)^k$. But with $\epsilon \ll 1/|S|$, most samples will contain only zeros, thus being in error by $(1/\epsilon)^k$. Or it does contain at least a value of $(1/\epsilon)^{k+1}$, then the estimated mean on the left is much larger than the "true" value. Examples like these show that there are (within a probability model) no consistent estimators, tests or confidence intervals of expectations when all probability models are taken into account, even within an i.i.d. setup.[22]

This may seem to contradict traditional wisdom from the statistical textbooks. But the textbooks rely on very narrow 'assumptions' that provide optimality results at fixed parameter points. Uniformity of results for all sets of parameters is rarely considered.

Motivated by similar examples, Davies (1995) argued to acknowledge the "approximate nature of probability models" by taking all probability models as adequate models of a data set if data simulated from the

[21] See Hennig (2007) for some further comments and Leeb/Pötscher (2006) on some general impossibility results for model selection and the distribution of estimators conditional on specification tests.

[22] See Bahadur/Savage (1956), Gleser/Hwang (1987), Lehmann/Loh (1990), Pfanzagl (1998), and Romano (2004) for further details.

models are "very much like the sample actually obtained".[23] If this is to work in the above example, one has to insist on a known bound on the support of the probability model. Otherwise, any value of an expectation (or any other moment) would be an 'adequate' model for any given data set.[24]

What is expressed by (2.24) then is not a truth deducible within the probability model, i.e. when the left hand side refers to a realisation of the model variables. But it is also obviously wrong when the left hand side is taken to be the empirical mean over some sample, the right hand side is interpreted as a mean over the population and nothing is 'assumed' about the values in the population.[25] In contrast to both positions, the one that makes (2.24) a consequence of the model and the one that naively sees it as an empirical justification of probability models, it seems better to see (2.24) as an expression of a further constitutive element of a particular type of a probabilistic model. It expresses the intent to treat empirical means to gauge the performance of subsets of probability models. It singles out a certain 'estimation principle' that can be used to criticise a given probabilistic model. But it in no way implies that it is the only way to do so. Nor does its acceptance imply that there is some 'true' model.

The CAR condition (2.21), on the other hand, can not be used in such a way simply because there is no empirical way to express it. It is also not constitutive in Matheron's sense. It singles out one particular set of solutions from all probability distributions based on a probability concept, at least when (2.24) is accepted. But it can not, as an 'assumption', justify that particular choice. Nor does it constitute the form of models or the set of tools that are to be used within a certain type of models. And it does not provide means to argue about the model. It rather provides a reference point for further speculations about the situation within a

[23] Note that this reverses the roles of probability models and observations as expressed by Nelson(2002) cited above.

[24] See also some further comments in section 4.

[25] It is at most a different figure of speech if 'population' here is interpreted within a probability model as a naive frequentist might suggest, whether directly referring to an 'infinite population' or via a super-population model. Since limiting frequencies do not express anything about any finite subsequence, counterexamples are easily produced.

probability model. Such speculations can not be completely arbitrary, they are at least constrained by the mathematical rules of probability. But they are clearly different from assumptions about reality as well.

2.2.2. CAR in Hierarchically Structured Models

Before looking into the use of such principled speculations I still have to show whether the CAR formulation (2.21) is general enough to deal with all types of incomplete data beyond the simple case of either completely missing or exact observations.

Suppose that there is some partial information on Y beyond that of Y being either completely missing ($Y^* = \mathcal{Y}$) or exactly given ($Y^* = \{y\}$) for some $y \in \mathcal{Y}$. Example 1 is a case in point since it differentiates between 'employed', 'unemployed', and 'out of labour force' (coded as $\mathcal{Y} = \{1, 2, 3\}$) and allows for partial answers: if a respondent does not differentiate between 'unemployed' and 'out of labour force', this answer is coded as $y^* = \{2, 3\}$. In this case, the information $y^* = \{2, 3\}$ cannot simply be ignored. But the CAR condition now says that the only information supplied by the respondent is that her employment status is in fact either 'unemployed' or 'out of labour force' and must be treated as such. There is then no need to contemplate why she has chosen to answer in the way she did, at least not as long as one is interested in the probabilities of 'employment', 'unemployment', and 'out of labour force'.

Using the formulation (2.21) in conjunction with (2.22) and the previously defined short hand notation, plugging in the values for $\Pr(Y^* = .)$ from Example 1, one arrives at:

$$\begin{aligned}
p_1 &= p_{\{1\}} + p_{\{1,2,3\}}p_1 = 0.5 + 0.2p_1 \\
p_2 &= p_{\{2\}} + p_{\{2,3\}}\frac{p_2}{p_2 + p_3} + p_{\{1,2,3\}}p_2 \\
&= 0.1 + 0.1\frac{p_2}{p_2 + p_3} + 0.2p_2 \\
p_3 &= p_{\{3\}} + p_{\{2,3\}}\frac{p_3}{p_2 + p_3} + p_{\{1,2,3\}}p_3
\end{aligned}$$

$$= 0.1 + 0.1\frac{p_3}{p_2 + p_3} + 0.2p_3$$

The first equation is easily solved, giving $p_1 = 5/8 = 0.625$. From this it follows that $p_2 + p_3 = 3/8$ so that

$$p_2 = 0.1 + \frac{0.1}{3/8}p_2 + 0.2p_2$$
$$p_3 = 0.1 + \frac{0.1}{3/8}p_3 + 0.2p_3$$

giving $p_2 = p_3 = 3/16 = 0.1875$.

This example, while exceedingly simple, is typical for hierarchically structured incomplete data. In this case, one can always proceed from top down through the hierarchy of partitions to provide simple linear equations for the probabilities of Y in subsets of $\mathcal{Y}^*$. To see this, note that for a subset $A \in \mathcal{Y}^*$

$$\begin{aligned}
\Pr(Y \in A) &= \sum_{y \in A} \Pr(Y = y) \\
&= \sum_{y \in A} \sum_{\substack{y^* \in \mathcal{Y}^* \\ y \in y^*}} \Pr(Y = y \mid Y^* = y^*) \Pr(Y^* = y^*) \\
&= \sum_{y \in A} \sum_{y^* \in \mathcal{Y}^*} \mathbb{1}[y^*](y) \Pr(Y = y \mid Y^* = y^*) \Pr(Y^* = y^*) \\
&= \sum_{y^* \in \mathcal{Y}^*} \sum_{y \in A} \mathbb{1}[y^*](y) \Pr(Y = y \mid Y \in y^*) \Pr(Y^* = y^*) \\
&= \sum_{y^* \in \mathcal{Y}^*} \Pr(Y \in A \cap y^* \mid Y \in y^*) \Pr(Y^* = y^*) \\
&= \sum_{\substack{y^* \in \mathcal{Y}^* \\ \Pr(Y^* = y^*) > 0}} \Pr(Y^* = y^*) \frac{\Pr(Y \in A \cap y^*)}{\Pr(Y \in y^*)} \qquad (2.25)
\end{aligned}$$

generalising (2.22). Since $\{Y^* = y^*\} \subseteq \{Y \in y^*\}$, the fractions in the last expression are well defined. Note that (2.25) can be written alternatively as

$$\Pr(Y \in A) = \sum_{\substack{y^* \in \mathcal{Y}^* \\ \Pr(Y^*=y^*)>0}} \Pr(Y^* = y^* \mid Y \in y^*) \Pr(Y \in A \cap y^*) \quad (2.26)$$

In the hierarchically structured case, i.e. when $\mathcal{Y}^*$ is the union of refinements of a partition, if A is an element of $\mathcal{Y}^*$ and if $y^* \cap A \neq \emptyset$ for some y^*, then either $y^* \subseteq A$ or $A \subseteq y^*$. Now if $y^* \subseteq A$, then the fraction on the right hand side is 1. That is, at most terms of a coarser partition than that to which A belongs contribute non-trivially to the sought for probabilities $\Pr(Y \in A)$.

Starting at the coarsest non-trivial partition (i.e. ignoring the uninformative level comprising all of $\mathcal{Y}$, the set indicating completely missing information) in $\mathcal{Y}^*$, there are subsets $A_1, \ldots, A_k \in \mathcal{Y}^*$ forming a partition of $\mathcal{Y}$. For these sets, (2.25) takes the form

$$\begin{aligned}\Pr(Y \in A_i) &= \sum_{\substack{y^* \in \mathcal{Y}^* \\ \Pr(Y^*=y^*)>0}} \Pr(Y^* = y^*) \frac{\Pr(Y \in A_i \cap y^*)}{\Pr(Y \in y^*)} \\ &= \Pr(Y \in A_i) \Pr(Y^* = \mathcal{Y}) + \sum_{y^* \subseteq A_i} \Pr(Y^* = y^*)\end{aligned}$$

resulting in the explicit solution

$$\Pr(Y \in A_i) = \frac{\sum_{y^* \subseteq A_i} \Pr(Y^* = y^*)}{\Pr(Y^* \neq \mathcal{Y})}$$

which is always defined unless $\Pr(Y^* = \mathcal{Y}) = 1$. But in the latter case, obviously no information about $\Pr(Y \in .)$ can be derived from the distribution of Y^*.

Proceeding down the hierarchy one level, one finds a refinement of the partition $\{A_1, \ldots, A_k\}$, say $\{A_{1_1}, \ldots, A_{1_{k1}}, \ldots, A_{k_1}, \ldots, A_{k_{kk}}\}$. But for this refinement, the denominators of the fractions $\Pr(Y \in A_{i_j} \cap y^*)/\Pr(Y \in y^*)$ are either known from the previous computation (when

$y^* \in \{A_1, \ldots, A_k, \mathcal{Y}\}$) or they equal 1 for subsets of the refinement. Thus,

$$\begin{aligned}
\Pr(Y \in A_{i_j}) &= \sum_{\substack{y^* \in \mathcal{Y}^* \\ \Pr(Y^* = y^*) > 0}} \Pr(Y^* = y^*) \frac{\Pr(Y \in A_{i_j} \cap y^*)}{\Pr(Y \in y^*)} \\
&= \Pr(Y \in A_{i_j}) \left(\Pr(Y^* = \mathcal{Y}) + \frac{\Pr(Y^* = A_i)}{\Pr(Y \in A_i)} \right) \\
&\quad + \sum_{y^* \subseteq A_{i_j}} \Pr(Y^* = y^*)
\end{aligned}$$

where it is assumed that $\Pr(Y \in A_i) > 0$ (otherwise, the term is simply dropped from the sum). But the denominator of the second term, $\Pr(Y \in A_i)$, is known from the previous step. Once again, the resulting simple linear equations can be solved directly and one may proceed further down to the finest partition.

Consequently, the CAR condition together with the consistency condition reduces to a sequence of equations that can be solved recursively. In particular a CAR model always exists and is unique down to the finest partition of $\mathcal{Y}$, i.e. the distribution of Y, $\Pr(Y \in .)$, can be determined for all sets in $\mathcal{Y}^*$. If the latter contains all singletons with positive probabilities, then the distribution of Y is uniquely identified from the coarsened data together with the CAR condition.

This will be true even when the information from the distribution of Y^* is extremely scarce. Suppose that $\Pr(Y^* = \mathcal{Y}) = 1 - \epsilon$ and $\Pr(Y^* = \{y_i\}) = \epsilon_i$ with $\sum_i \epsilon_i = \epsilon$ and $\epsilon, \epsilon_i > 0$ for some small ϵ. Then the set of consistent distributions of Y is a simplex whose volume is arbitrarily close to the full probability simplex. In other words, the consistency conditions do not restrict the set of consistent probabilities beyond their being positive. But even in this case, the CAR condition leads to a unique result. The effectiveness of the CAR condition to single out one particular distribution from the many possible ones simply does not depend on the informativeness of the distribution of Y^*.

A noteworthy further consequence of the CAR condition in the hierarchical case is that it leads to a rational solution when the known $\Pr(Y^* = y^*)$

are rational. Thus, if the probabilities of the coarsened sets are taken to be relative frequencies, the CAR solution can be interpreted as a relative frequency as well. In particular, the construction does not necessarily lead beyond an interpretation based on the classical sampling theory.

From the CAR condition (2.21) and the marginal distribution of Y^* one can also write down the joint distribution of (Y, Y^*). The joined density follows from

$$\begin{aligned}\Pr(Y = y, Y^* = y^*) &= \Pr(Y = y \mid Y \in y^*) \Pr(Y^* = y^*)\\ &= \mathbb{1}[y^*](y) \Pr(Y^* = y^*) \frac{\Pr(Y = y)}{\sum_{y \in y^*} \Pr(Y = y)}\end{aligned} \tag{2.27}$$

for $\Pr(Y^* = y^*) > 0$. Note that the consistency condition (2.3) implies that the denominator in the fraction above is > 0. Furthermore, $\Pr(Y = y, Y^* = y^*) = 0$ for $\Pr(Y^* = y^*) = 0$. Note also that this is an allocation of the form given in (2.19).

In Example 1, the joined density is readily seen to be

	$\{1\}$	$\{2\}$	$\{3\}$	$\{2,3\}$	$\{1,2,3\}$	
1	1/2	0	0	0	1/8	5/8
2	0	1/10	0	1/20	3/80	3/16
3	0	0	1/10	1/20	3/80	3/16
	1/2	1/10	1/10	1/10	2/10	

2.2.3. Non-Hierarchical CAR

Suppose next that $\mathcal{Y}^*$ is no longer hierarchical. A simple example is given by my Example 2 where

$$\begin{aligned}&\mathcal{Y}_2^* = \{\{1\}, \{2\}, \{3\}, \{1,2\}, \{2,3\}, \{1,2,3\}\}\\ &p_{\{1\}} = 0.5, p_{\{2\}} = 0.1, p_{\{3\}} = 0.1, p_{\{1,2\}} = 0.1, p_{\{2,3\}} = 0.1,\\ &p_{\{1,2,3\}} = 0.1\end{aligned}$$

Here the equations (2.22) become

$$
\begin{aligned}
p_1 &= 0.5 + 0.1\frac{p_1}{p_1 + p_2} + 0.1p_1 \\
p_2 &= 0.1 + 0.1\frac{p_2}{p_1 + p_2} + 0.1\frac{p_2}{p_2 + p_3} + 0.1p_2 \\
p_3 &= 0.1 + 0.1\frac{p_3}{p_2 + p_3} + 0.1p_3
\end{aligned}
$$

where the restriction $p_1+p_2+p_3 = 1$ was used in the last terms. These equations are no longer linear. Multiplying through by p_1+p_2, $(p_1+p_2)(p_2+p_3)$, and $p_2 + p_3$, one is led to a system of polynomial equations which in this case simplifies because the first and last equations depend only on p_1, p_2 and p_2, p_3, respectively. In fact, the system can be triangularised resulting in[26]

$$
\begin{aligned}
243p_3^3 - 261p_3^2 + 73p_3 - 6 &= 0 \\
7p_2 - 27p_3^2 + 33p_3 - 6 &= 0 \\
7p_1 - 162p_3^2 + 135p_3 - 22 &= 0
\end{aligned}
$$

The first equation has the three real solutions $p_3 = 2/9$ or $p_3 = (23 \pm \sqrt{205})/54$. Plugging the rational solution 2/9 into the second and third equation forces $p_1 = p_2 = 0$ so that this is no probability distribution. Using $(23 + \sqrt{205})/54$ gives a negative solution in the second equation. Only the last solution, which obviously is irrational, provides a solution that also respects the restrictions of a probability model. Thus, the solution is still unique and is given by $p_1 = (95 - \sqrt{205})/126 \approx 0.64033$, $p_2 = (5\sqrt{205} - 34)/189 \approx 0.19888$, $p_3 = (23 - \sqrt{205})/54 \approx 0.16078$.

[26] There is a further triangularisation which, however, violates the requirement $p_1 + p_2 + p_3 = 1$. The solution was computed using the Singular package (http://www.singular.uni-kl.de) using lexicographical ordering of the monomial terms. Any other software that allows to compute Gröbner bases and resultants can obviously be used as well. Further pertinent free software packages for such computations include Macaulay 2 (http://www.math.uiuc.edu/Macaulay2/) and CoCoA (http://cocoa.dima.unige.it/). Another program that uses homotopy methods is PHCpack (http://www.math.uic.edu/~jan/download.html).

The elements of the solution vector, however, are no longer rational numbers.[27] Consequently, there is no sample size and there are no relative frequencies such that the CAR requirement (2.21) holds exactly, whatever the number of observations.

If the CAR condition is supposed to hold, however, the implied joined density of (Y, Y^*) is given (using (2.27)) by

	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,2\}$	$\{2,3\}$	$\{1,2,3\}$	
1	1/2	0	0	0.0763	0	0.0640	0.6403
2	0	1/10	0	0.0237	0.0553	0.0199	0.1989
3	0	0	1/10	0	0.0447	0.0161	0.1608
	1/2	1/10	1/10	1/10	1/10	1/10	

In the case of Example 3, where

$$\mathcal{Y}_3^* = \{\{1\}, \{2\}, \{1,2\}, \{2,3\}, \{1,2,3\}\}$$
$$p_{\{1\}} = 0.5, p_{\{2\}} = 0.1, p_{\{1,2\}} = 0.1, p_{\{2,3\}} = 0.1, p_{\{1,2,3\}} = 0.2$$

the condition (2.22) gives

$$\begin{aligned} p_1 &= 0.5 + 0.1\frac{p_1}{p_1+p_2} + 0.2p_1 \\ p_2 &= 0.1 + 0.1\frac{p_2}{p_1+p_2} + 0.1\frac{p_2}{p_2+p_3} + 0.2p_2 \\ p_3 &= 0.1\frac{p_3}{p_2+p_3} + 0.2p_3 \end{aligned}$$

If one adds the requirement $p_1 + p_2 + p_3 = 1$ to the equations, the triangularised set of equations becomes

$$\begin{aligned} 16p_3^2 - 9p_3 &= 0 \\ 7p_2 + 9p_3 - 2 &= 0 \end{aligned}$$

[27] Whether there is an algorithm to decide whether a given system of polynomials with rational coefficients has rational solutions is Hilbert's 10th problem. Such an algorithm does not exist for integer solution. Whether there is one for rational solutions is still unknown.

$$p_1 + p_2 + p_3 - 1 = 0$$

where the two real solutions to the first equation are $p_3 = 0$ and $p_3 = 9/16$. The second solution leads to a negative value for p_2. Thus the solution is given by $p_1 = 5/7 = 0.71429, p_2 = 2/7 = 0.28571, p_3 = 0$ which is now on the boundary of the distributions compatible with the consistency constraints, though again rational.

The joined density of (Y, Y^*) for Example 3 is given by

	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,2\}$	$\{2,3\}$	$\{1,2,3\}$	
1	1/2	0	0	1/14	0	2/14	5/7
2	0	1/10	0	1/35	1/10	2/35	2/7
3	0	0	0	0	0	0	0
	1/2	1/10	0	1/10	1/10	2/10	

In general, non-hierarchical coarsening patterns will lead to polynomial equations for the probabilities of interest. One may try to gain insight in the solutions of such systems by using the methods of Sturmfels (2002) and Cox et al. (2002).[28] An interesting method that uses only the structure of $\mathcal{Y}^*$ to bound the number of solutions from above is related to the (mixed) volume of certain polytopes.[29]

As far as I can see, the bound is of no immediate help for the solution of CAR equations. But the polytopes provide another graphical method represent the structure of $\mathcal{Y}^*$. To illustrate, one can write out the system of polynomial equations for Example 2, this time without using $p_1+p_2+p_3 = 1$ and multiplying with all denominators. This leads to the following three polynomials:

$$p_1^3 + 2p_1^2p_2 + p_1^2p_3 - 0.7p_1^2 + p_1p_2p_3 + p_1p_2^2 - 1.2p_1p_2 - 0.6p_1p_3$$
$$- 0.5p_2^2 - 0.5p_2p_3$$
$$p_2^4 + 2p_2^3p_3 + 2p_1p_2^3 - 0.2p_2^3 + p_2^2p_3^2 + 3p_1p_2^2p_3 - 0.6p_2^2p_3 + p_1^2p_2^2$$

[28] A short introduction with a view towards applications in statistics is Pistone et al. (2001).

[29] Rojas (2003) provides an accessible introduction.

$$
\begin{aligned}
&- 0.6p_1p_2^2 + p_1p_2p_3^2 - 0.2p_2p_3^2 + p_1^2p_2p_3 - 0.6p_1p_2p_3 - 0.2p_1^2p_2 \\
&- 0.1p_1p_3^2 - 0.1p_1^2p_3 \\
p_3^3 &+ 2p_2p_3^2 + p_1p_3^2 - 0.1p_3^2 + p_2^2p_3 + p_1p_2p_3 - 0.2p_2p_3 - 0.2p_1p_3 \\
&- 0.1p_2^2 - 0.1p_1p_2
\end{aligned}
$$

The exponents of the monomial terms of the polynomials encode to how many of the sets in $\mathcal{Y}^*$ a particular y belongs. They also encode how often a pair (y, y') is a subset of the sets in $\mathcal{Y}^*$ etc. Writing down the exponents as vectors of integers in the order p_1, p_2, p_3 gives the following three sets of points:

$$
\begin{aligned}
A_1 = \{&(3,0,0),(2,1,0),(2,0,1),(2,0,0),(1,1,1),(1,2,0),\\
&(1,1,0),(1,0,1),(0,2,0),(0,1,1)\}\\
A_2 = \{&(0,4,0),(0,3,1),(1,3,0),(0,3,0),(0,2,2),(1,2,1),\\
&(0,2,1),(2,2,0),(1,2,0),(1,1,2),(0,1,2),(2,1,1),\\
&(1,1,1),(2,1,0),(1,0,2),(2,0,1)\}\\
A_3 = \{&(0,0,3),(0,1,2),(1,0,2),(0,0,2),(0,2,1),(1,1,1),\\
&(0,1,1),(1,0,1),(0,2,0),(1,1,0)\}
\end{aligned}
$$

Taken as points in $\mathbb{R}^3$ with integer coordinates, one may construct the convex hull of the points in A_1, A_2, A_3. These are the *Newton* polytopes of the respective polynomials. Now the Newton polytope of the product of polynomials is just the Minkowski sum of the Newton polytopes. Thus in so far as the Newton polytope of a polynomial encodes information about the zeros of the polynomial, all Minkowski sums that can be constructed from the three sets should give information on the zeros of the system of equations. Figure 2.9 demonstrates the constructions.[30]

[30] The mixed volume in this case is defined as $\text{MV}(A_1, A_2, A_3) = \text{Vol}(\text{conv}(A_1) + \text{conv}(A_2) + \text{conv}(A_3)) - \text{Vol}(\text{conv}(A_1) + \text{conv}(A_2)) - \text{Vol}(\text{conv}(A_1) + \text{conv}(A_3)) - \text{Vol}(\text{conv}(A_2) + \text{conv}(A_3)) + \text{Vol}(\text{conv}(A_1)) + \text{Vol}(\text{conv}(A_2)) + \text{Vol}(\text{conv}(A_3))$. A multiple of this number bounds the number of the non-zero isolated complex solutions to the system of polynomial equations. But what is needed here, of course, is a bound on the number of real solutions in the probability simplex. It is still unclear whether the geometric methods can provide insight into this problem.

The actual solution to the CAR equations, however, depends not only on the structure of the set $\mathcal{Y}^*$ but also on the coefficients $\Pr(Y^* = y^*)$. It is, moreover, not obvious how to include the inequalities of the consistency requirements into the framework of systems of polynomial equations. One may add slack variables to represent the constraints. But as far as I can judge from simple examples, such a move tends to complicate matters in many cases without any real gain of insight.

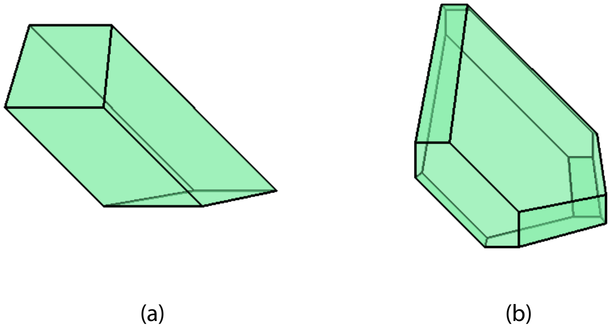


Figure 2.9.: The Newton polytope $\operatorname{conv}(A_1)$ and the Minkowski sum $\operatorname{conv}(A_1) + \operatorname{conv}(A_2) + \operatorname{conv}(A_3)$ from Example 2.

2.2.4. CAR is Everything

Numerical solutions are much easier (though less explicitly) computed using (2.22) directly. After all, that equation expresses a fixed point property which may be used as a recipe to compute a stationary point satisfying both the consistency requirements and the additional CAR condition. Starting from any assignment of positive values to $\Pr(Y \in .)$ one can use the right hand side of (2.22) to update the probability so that the consistency constraints are always satisfied. The procedure can be iterated and the solutions will converge to a fixed point of the equations

(2.22) if at least one exists.[31]

In section 2.1.6 it was shown that any choice of a probability measure with $\Pr(Y = y) > 0$ for all y will give rise to an allocation respecting the consistency requirements via (2.22). In fact, the identity (2.22) is just the marginal distribution of the selector corresponding to the allocation (2.19). Starting values for iterating (2.22) can therefore be chosen more generally from all selectors of allocations.

Note that the equations (2.22) are the self-consistency equations introduced by Efron (1967) and discussed by Tsai and Crowley (1985).[32] Since

$$\Pr(Y = y) = \mathbb{E}_{Y^*}\left(\mathbb{E}_{Y|Y^*}\left(\mathbb{1}[Y = y] \mid Y^*\right)\right)$$

by (2.21) one gets

$$\begin{aligned}\Pr(Y = y) &= \mathbb{E}_{Y^*}\left(\Pr(Y = y \mid Y^*)\right)\\ &= \mathbb{E}_{Y^*}\left(\Pr(Y = y \mid Y \in Y^*)\right)\\ &= \sum_{\substack{y^* \in \mathcal{Y}^*\\ y^* \ni y\\ \Pr(Y^* = y^*) > 0}} \Pr(Y^* = y^*)\frac{\Pr(Y = y)}{\Pr(Y \in y^*)}\end{aligned}$$

Turning to the existence of a solution, note that the right hand side of (2.22) is a continuous function of $\Pr(Y = y)$ since the denominators in the equation are strictly positive and will stay so by the consistency requirement. Moreover, the image under this map of any consistent probability on $\mathcal{Y}$ (an element of the compact convex set of consistent

[31] Numerical applications of the algorithm provide only approximations to actual fixed points. The numerical solutions may not even be close to actual solutions of (2.22). As usual, results must be checked by other means.

[32] Orchard/Woodbury (1972) termed the construction principle the 'missing information principle'. A general version is discussed by Lai/Ying (1994). The equations are a special case of the (non-parametric) EM-algorithm. Wu (1983) discusses convergence properties. Tsodikov (2003) extends the idea to general semi-parametric models, Pons (2006) discusses applications to Markov processes. Subramanian (2003) uses the missing information principle in the case of general censored regression models where the censoring depends on covariates.

probabilities) is necessarily again an element of the set of consistent probabilities. Therefore, there always is at least one solution to the equations (2.22).[33]

Because there is always a solution satisfying CAR for any given distribution on $\mathcal{Y}^*$, CAR can not be ruled out empirically. Gill et al. (1997) summarised the finding with the slogan 'CAR is everything'. It is always possible to take the coarsened data at face value in the sense of the CAR condition (2.21). Put still differently, no distribution on $\mathcal{Y}^*$ will contradict the CAR condition and, conversely, the CAR condition does not restrict the distribution of Y^*.[34]

The only caveat so far was that CAR may result in irrational probabilities for some of the y, even if all $\Pr(Y^* = y^*)$ are rational. In this sense, the "simple device" of Rubin and Little takes one beyond the strict limits of a finite sample approach.

2.2.5. Uniqueness of CAR

Certainly, solutions to the equations will not be unique in general. In particular, a distribution that puts zero probability on some $y \in \mathcal{Y}$ but having $\Pr(Y \in y^*) > 0$ for all y^* with $\Pr(Y^* = y^*) > 0$ may satisfy the equations as well as some other distribution with $\Pr(Y = y) > 0$. This will certainly happen when the partition of $\mathcal{Y}$ induced by the sets in $\mathcal{Y}^*$ does contain a set, say A, with cardinality $|A| > 1$. If $y, y' \in A$, then the sets in $\mathcal{Y}^*$ do not allow to distinguish between y and y'. In such a case, it can not be expected that CAR by itself identifies the probabilities of y and y'. Modifying Example 1 by eliminating all singletons except $\{1\}$ from $\mathcal{Y}^*$ gives

Example 1a

$$\mathcal{Y}^*_{1a} = \{\{1\}, \{2, 3\}, \{1, 2, 3\}\}$$

[33] This is the mean value theorem for $|\mathcal{Y}| = 2$ and Brouwer's fixed point theorem for general (finite) $|\mathcal{Y}|$.

[34] Molenberghs et al. (2008) formulate the equivalent result in a parametric setting, assuming parameter distinctness.

$$p_{\{1\}} = 0.5, p_{\{2,3\}} = 0.3, p_{\{1,2,3\}} = 0.2$$

say. The equations (2.22) become

$$\begin{aligned} p_1 &= 0.5 + 0.2p_1 \\ p_2 &= 0.3\frac{p_2}{p_2 + p_3} + 0.2p_2 \\ p_3 &= 0.3\frac{p_3}{p_2 + p_3} + 0.2p_3 \end{aligned}$$

whose solution set is given by $p_1 = 5/8$ and any partition of the remaining mass of 3/8 between p_2 and p_3. This is a closed, connected and convex subset of the probability simplex. The set of solutions within the set of consistent distributions is shown in Figure 2.12.

Restricting for the time being $\mathcal{Y}^*$ to those elements y^* such that $\Pr(Y^* = y^*) > 0$, a very special case arises when $\mathcal{Y}^*$ is just a partition of $\mathcal{Y}$. That happens when only simply grouped data are available. This is a hierarchically ordered model and thus $\Pr(Y \in y^*)$ is uniquely determined for all $y^* \in \mathcal{Y}^*$. In particular, $\Pr(Y \in y^*) = \Pr(Y^* = y^*)$. But for y^* with $|y^*| > 1$, this is the only restriction on the probabilities $\Pr(Y = y)$ for $y \in y^*$, i.e. the probabilities can vary between 0 and $\Pr(Y \in y^*)$. In this situation, the set of consistent distributions coincides with the set of CAR distributions. To see this, suppose that $\Pr(Y \in .)$ is a consistent distribution. Then $\{Y^* = y^*\} \subseteq \{Y \in y^*\}$ by consistency and $\{Y \in y^*\} \subseteq \{Y^* = y^*\}$ by construction. Thus, the CAR condition (2.21) follows. This is the only situation where the CAR condition is implied by the structure of $(\mathcal{Y}, \mathcal{Y}^*)$ and is therefore always justified. There simply are no distributions consistent with the structure that could contradict CAR.[35] Furthermore, in this case the probability distribution of Y is determined on the σ-algebra generated by the partition.

Gill et al. (1997: 262) claimed that the only form of non-uniqueness is the one exhibited by Example 1a, and that in particular the solution to the CAR conditions are always unique on sets y^* with $\Pr(Y^* = y^*) > 0$. This is obviously the most one can hope for. Suppose now that $\Pr(Y \in y^*)$ is

[35] See the discussion by Grünwald/Halpern (2003: Proposition 4.1) and De Cooman/Zaffalon (2004: Sect. 3).

in fact uniquely identified from the distribution of Y^*. Then also the probability of complements, $\Pr(Y \in y^{*c})$, is identified. In other words, the probabilities of subsets $\{y^{*\prime} \in \mathcal{P}(\mathcal{Y}) \,|\, y^{*\prime} \in \mathcal{Y}^* \cup y^{*\prime c} \in \mathcal{Y}^*\}$ are identified. Furthermore, disjoint unions and set differences $y^* \setminus y^{*\prime}$ for $y^{*\prime} \subseteq y^*$ are identified. This set of subsets of $\mathcal{Y}$ will be denoted by $\mathcal{D}(\mathcal{Y}^*)$.[36]

In the hierarchically ordered case, this implies that the probabilities of the finest partition (and their unions) are known. Since the elements of the finest partition together with all their unions forms a σ-algebra, as is the case for simply grouped data, $\mathcal{D}(\mathcal{Y}^*) = \sigma(\mathcal{Y}^*)$.

In all other cases, however, the sets $y^{*\prime}$ for which $\Pr(Y \in y^{*\prime})$ would be identified from the uniqueness of CAR for $\Pr(Y^* = y^*) > 0$ will not form a σ-algebra. Consider Example 5 below:

Example 5

Suppose $\mathcal{Y} = \{1, 2, 3, 4\}$ and

$$\mathcal{Y}_5^* = \{\{1, 2\}, \{1, 3\}, \{2, 4\}, \{3, 4\}\}$$
$$p_{\{1,2\}} = p_{\{1,3\}} = p_{\{2,4\}} = p_{\{3,4\}} = 1/4$$

The set of solutions of the CAR equations (2.22) is given by $p_1, 1/2 - p_1, 1/2 - p_1, p_1$ for $p_1 \in [0, 1/2]$, a line segment joining $(0, 1/2, 1/2, 0)$ to $(1/2, 0, 0, 1/2)$. This is a closed, convex set. Moreover, $\Pr(Y \in \{1, 2\}) = 1/2 = \Pr(Y \in \{3, 4\}) = \Pr(Y \in \{2, 4\}) = \Pr(Y \in \{1, 3\})$ as suggested by the Gill et al. conjecture.

The hypergraph corresponding to this example is given in the following Figure 2.10. Its symmetry suggests why the the probability of the above mentioned sets is identified from the distribution of Y^* together with the CAR condition.

[36] The elements of $\mathcal{D}(\mathcal{Y}^*)$ form what is known as a λ-system in measure theory, a set of subsets of $\mathcal{Y}$ containing the empty set and being closed under the formation of complements and disjoint unions. See Pollard (2002: 42). Note that Grünwald and Halpern (2003: 251) included also intersections into what they termed $\mathcal{R}$-atoms, the building blocks for their algebraic characterisation of CAR.

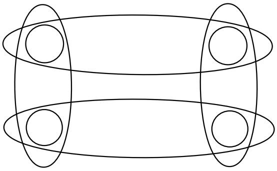


Figure 2.10.: Hypergraph for Example 5. The innermost circles represent the sets $\{1,2\}$, $\{1,3\}$, $\{2,4\}$, and $\{3,4\}$. The outer ellipses represent the elements 1, 2, 3, 4.

Looking at the structure of subsets whose probability is uniquely determined by the CAR condition, neither $\Pr(Y = 1)$ nor $\Pr(Y \in \{1,2,3\})$ are identified unless one of the $\Pr(Y^* = .)$ equals 0. But $\{1\} = \{1,2\} \cap \{1,3\}$ so that $\{1\}$ is an element of the σ-algebra generated by $\{\{1,2\}, \{1,3\}, \{2,4\}, \{3,4\}\}$. Similarly, $\{1,2,3\} = \{1,2\} \cup \{1,3\}$ is an element of the σ-algebra but its probability is not identified. The identified probabilities can not be extended to the σ-algebra because the unions and intersections here are taken among non-nested elements of $\mathcal{Y}^*$ and such that these unions and intersections are not themselves elements of $\mathcal{Y}^*$. Therefore, in this example $\mathcal{D}(\mathcal{Y}^*) \subsetneq \sigma(\mathcal{Y}^*)$.

I have already shown that if $\mathcal{Y}^*$ is hierarchically ordered, then necessarily $\mathcal{D}(\mathcal{Y}^*) = \sigma(\mathcal{Y}^*)$. Conversely, suppose that $\mathcal{D}(\mathcal{Y}^*) = \sigma(\mathcal{Y}^*)$ and let $y^{*\prime}$ be an arbitrary element of $\sigma(\mathcal{Y}^*)$. Being an element of $\sigma(\mathcal{Y}^*)$, $y^{*\prime}$ is the union of elements of $\mathcal{Y}^*$ or their complements, or the complement of such a union. Write $y^{*\prime} = \cup y^{*\prime\prime}$ with either $y^{*\prime\prime} \in \mathcal{Y}^*$ or $y^{*\prime\prime c} \in \mathcal{Y}^*$. For this to be also an element of $\mathcal{D}(\mathcal{Y}^*)$, any pair of sets $y_1^{*\prime\prime}$ and $y_2^{*\prime\prime}$ must satisfy either $y_1^{*\prime\prime} \subseteq y_2^{*\prime\prime}$ or $y_2^{*\prime\prime} \subseteq y_1^{*\prime\prime}$ or $y_1^{*\prime\prime} \cap y_2^{*\prime\prime} = \emptyset$ so that $\sigma(\mathcal{Y}^*)$ is itself hierarchically ordered. If this holds true for $\sigma(\mathcal{Y}^*)$, it must be true for $\mathcal{Y}^*$, a subset of $\sigma(\mathcal{Y}^*)$. Thus, $\mathcal{Y}^*$ is hierarchically ordered. This clarifies the relation between the structure of the set $\mathcal{D}(\mathcal{Y}^*)$ (which according to the Gill et al. conjecture would be the largest set of subsets whose probabilities are uniquely determined by CAR) and the structure of $\mathcal{Y}^*$.

But is the form of non-uniqueness exemplified by Example 1a and Example 5 really the only one that is to be expected? Do we only have to care about non-uniqueness on sets of the above form? A variation of Example 3 shows that the claim of Gill et al. is false, at least when (2.21) is used as a definition of CAR.

In Example 3 it is the fact that $\Pr(Y^* = \{2\}) > 0$ from which $\Pr(Y = 2) > 0$ follows. But this implies $\Pr(Y = 3) = 0$. If the example is modified such that $\Pr(Y^* = \{2\}) = 0$, then it may seem to be difficult to identify separately the probabilities of $\{Y = 2\}$ and $\{Y = 3\}$. On the other hand, since $\Pr(Y^* = \{1\}) > 0$, $\Pr(Y = 1)$ should be uniquely identified. And from $\Pr(Y^* = \{1, 2\}) > 0$, $\Pr(Y \in \{1, 2\})$ should be identified. Therefore $\Pr(Y = 2) = \Pr(Y \in \{1, 2\} \setminus \{1\}) = \Pr(Y \in \{1, 2\}) - \Pr(Y = 1)$ is identified together with $\Pr(Y = 3)$. Thus, $\Pr(Y = .)$ is completely identified. In fact, CAR in that case implied a unique solution for the distribution of Y.

Now suppose that Example 3 is replaced by

Example 3a

$$\mathcal{Y}^*_{3a} = \{\{1\}, \{1, 2\}, \{2, 3\}, \{1, 2, 3\}\}$$
$$p_{\{1\}} = 0.5, p_{\{1,2\}} = 0.1, p_{\{2,3\}} = 0.1, p_{\{1,2,3\}} = 0.3$$

As in Example 3, since $\{2\} = \{1, 2\} \setminus \{1\}$, all the singletons should have unique probability and thus $\mathcal{D}(\mathcal{Y}^*) = \mathcal{P}(\mathcal{Y})$ when the Gill et al. conjecture was true. The probability distribution of Y would be uniquely determined. This was the case in Example 3.

The hypergraph of the modified example is given in Figure 2.11. It shows that the edges representing $\{Y = 2\}$ and $\{Y = 3\}$ are properly nested. This is already the case in Example 3 (see Figure 2.3) so that there is no reason to expect additional trouble.

The CAR condition (2.22) gives

$$p_1 = 0.5 + 0.1\frac{p_1}{p_1 + p_2} + 0.3p_1$$
$$p_2 = 0.1\frac{p_2}{p_1 + p_2} + 0.1\frac{p_2}{p_2 + p_3} + 0.3p_2$$

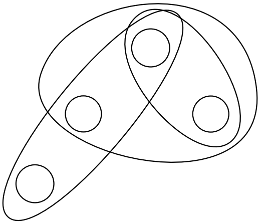


Figure 2.11.: Hypergraph for Example 3a. The innermost circles represent the sets $\{1\}, \{1,2\}, \{2,3\}$, and $\{1,2,3\}$. The outer ellipses represent the elements 1,2, and 3.

$$p_3 = 0.1\frac{p_3}{p_2 + p_3} + 0.3p_3$$

But now there are two different solutions to the equations:

$$p_1 = 5/6, p_2 = 1/6, p_3 = 0 \text{ and } p_1 = 6/7, p_2 = 0, p_3 = 1/7$$

The implied joined densities are

	$\{1\}$	$\{1,2\}$	$\{2,3\}$	$\{1,2,3\}$	
1	1/2	5/60	0	15/60	5/6
2	0	1/60	6/60	3/60	1/6
3	0	0	0	0	0
	1/2	1/10	1/10	3/10	

and

	$\{1\}$	$\{1,2\}$	$\{2,3\}$	$\{1,2,3\}$	
1	1/2	7/70	0	18/70	6/7
2	0	0	0	0	0
3	0	0	7/70	3/70	1/7
	1/2	1/10	1/10	3/10	

so that it is easy to check that both solutions indeed satisfy (2.21).

But now, contrary to the case of Example 3 and the conjecture by Gill et al. there is no proper subset of $\mathcal{Y}$ whose probability is uniquely determined by CAR. None of the singletons nor any of the two-element subsets of $\mathcal{Y}$ have identical probabilities under both solutions. Thus, the only subsets whose probabilities are uniquely determined are trivially $\emptyset$ and $\mathcal{Y}$. On the other hand, $\mathcal{D}(\mathcal{Y}^*) = \mathcal{P}(\mathcal{Y})$ so that the probability of Y ought to be uniquely identified if the conjecture was correct. Thus either the conjecture is false or one has to use a stronger condition than the CAR condition (2.21) to further reduce the solution set.

The same phenomenon as in the previous example occurs in the Monty Hall problem. If one puts $\Pr(Y^* = \{1, 2\}) = \Pr(Y^* = \{2, 3\}) = 1/2$, then the CAR equations result in

$$\begin{aligned}
p_1 &= 0.5\frac{p_1}{p_1 + p_2} \\
p_2 &= 0.5\frac{p_2}{p_1 + p_2} + 0.5\frac{p_2}{p_2 + p_3} \\
p_3 &= 0.5\frac{p_3}{p_2 + p_3}
\end{aligned}$$

Again, there are two different solutions to the equations:

$$p_1 = 0, p_2 = 1, p_3 = 0 \text{ and } p_1 = 1/2, p_2 = 0, p_3 = 1/2$$

Note that only the second solution depends on the probability distribution of Y^*: if one sets $p_{\{1,2\}} \coloneqq \Pr(Y^* = \{1, 2\})$ arbitrarily, then the second solution becomes $p_1 = p_{\{1,2\}}, p_2 = 0, p_3 = 1 - p_{\{1,2\}}$. The joined distributions are given by

	$\{1,2\}$	$\{2,3\}$	
1	0	0	0
2	1/2	1/2	1
3	0	0	0
	1/2	1/2	

	$\{1,2\}$	$\{2,3\}$	
1	1/2	0	1/2
2	0	0	0
3	0	1/2	1/2
	1/2	1/2	

Once again, condition (2.21) is easily verified.

The solutions to the CAR equations of the Examples 1a, 3a, and the Monty Hall solutions are depicted within the consistency region of the probability simplexes in Figure 2.12.

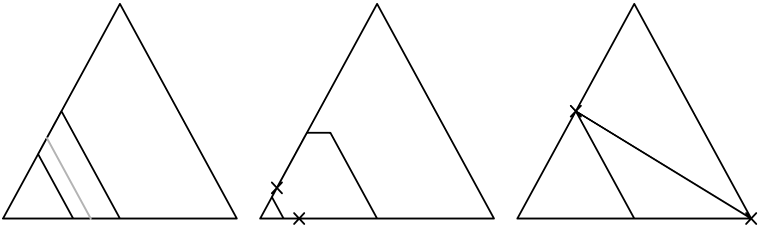


Figure 2.12.: The consistency polytopes and the solutions of the CAR equations for Examples 1a, 3a, and the Monty Hall problem. CAR solutions are indicated by × in the latter two cases.

But by the Gill et al. conjecture, since $\Pr(Y^* = \{1, 2\}) > 0$ identification of $\Pr(Y \in \{1, 2\})$ and therefore of $\Pr(Y = 3)$ should follow. Similarly, $\Pr(Y = 1)$ should be identified from $\Pr(Y^* = \{2, 3\}) > 0$. Thus $\Pr(Y = .)$ should be completely identified. Once again, however, the only sets with identical probabilities for both solutions are trivially $\emptyset$ and $\mathcal{Y}$.

What are the difficulties encountered when there are CAR solutions that differ on $\mathcal{D}(\mathcal{Y}^*)$? Are these difficulties related to the definition of CAR or to the rather optimistic conjecture of Gill et al.?

It is instructive to look at the Monty Hall problem in some more detail. I have argued in section 2.1.2 rather intuitively that there can not be a CAR solution in that example or in any example with the same hypergraph structure. Moreover, the two solutions presented above using the CAR condition (2.21) look strange in that they force zero probabilities on some alternatives that on a priori grounds would be taken to be possible. Even though the two solutions are contradictory, any one of them rules out a possibility solely based on a theoretical condition which therefore must be termed dogmatic.

Furthermore, the non-uniqueness is not of the kind that is necessarily present in all incomplete data situations. It does not arise from asking more than can reasonably be expected. If $\Pr(Y \in y^*)$ for both $y^* = \{1, 2\}$ and $y^* = \{2, 3\}$ were identified, then the distribution of Y would be identified as well. And as long as $\Pr(Y^* = \{1, 2\}) \notin \{0, 1\}$, all distributions of Y^* result in two distinct solutions. The difficulty is not related to the distribution of Y^*, but to the structure of $(\mathcal{Y}, \mathcal{Y}^*)$. The result does more to undermine the CAR condition as formulated in (2.21) than to question the reasonable expectation of Gill et al.

If these considerations are reasonable, one would need a more restrictive CAR formulation so that not all of the solutions consistent with (2.21) would be permitted under the label of CAR. This would rule out counter-intuitive results as the ones presented above.

But it must then also be allowed that no CAR solution exists. This would contradict the slogan that CAR is everything. Even more importantly, it would contradict the general believe, reiterated in nearly all expositions of the "simple device", that CAR is an 'untestable assumption'. In particular, the Monty Hall problem would not have a CAR solution and this can be asserted even without looking at the actual data.[37]

If despite all the difficulties CAR in the formulation (2.21) was adopted, the slogan 'Car is everything' would be vindicated and it would indeed be an 'untestable assumption'. But one then must be prepared to deal with nasty complications. CAR could be compatible with contradictory probabilistic accounts involving in extreme cases all non-trivial sets of the form $\{Y \in y^*\}$ for $y^* \in \mathcal{Y}^*$ simultaneously. And even in less extreme cases, elements of $\mathcal{D}(\mathcal{Y}^*)$ may not be unequivocally identified by CAR alone. Consequently, the concept would be of only limited use for either frequentist inference or Bayesian kinematics.

2.2.6. Further CAR Formulations

The CAR condition (2.21) can be reformulated in a number of useful ways that will also shed more light on the difficulties encountered in the previous section. One possible reformulation reverses the roles of

[37] The only prerequisite is that the set $\mathcal{Y}^*$ contains only elements with positive probability.

Y and Y^* in the conditional probabilities and is thus very helpful as a complement to (2.21). It can be written as

$$\Pr(Y^* = y^* \mid Y = y) \text{ is constant on } y \in y^* \tag{2.28}$$

for all $y^* \in \mathcal{Y}^*$ and all $y \in y^*$ such that $\Pr(Y = y) > 0$. Looking at the problem from this direction is somewhat closer to the emergence of incomplete data. The employment status of the respondents and possibly many other characterisations of the respondent and the interview situation are the background against which the respondents decide to participate in the survey and to answer its questions. The conditional probability of a particular response given (a subset) of the background variables may be used to model this. In particular, (2.28) says that the conditional probability of refusing to answer the question of employment status ($Y^* = \{1, 2, 3\}$ in Example 1) does not depend on that status. An immediate consequence is that

$$\Pr(Y^* = y^* \mid Y = y) = \Pr(Y^* = y^* \mid Y \in y^*) \tag{2.29}$$

for all $y^* \in \mathcal{Y}^*$ and all $y \in y^*$ such that $\Pr(Y = y) > 0$, a further form of the condition that is often useful.

Obviously, if (2.29) holds, then $\Pr(Y^* = y^* \mid Y = y)$ is constant for all $y \in y^*$ and $\Pr(Y = y) > 0$ so that (2.28) follows. In the other direction, fix an $y \in y^*$ with $\Pr(Y = y) > 0$ and suppose that (2.28) holds. Then

$$\begin{aligned}\Pr(Y^* = y^* \mid Y \in y^*) &= \frac{\sum_{y' \in y^*} \Pr(Y^* = y^* \mid Y = y') \Pr(Y = y')}{\sum_{y' \in y^*} \Pr(Y = y')} \\ &= \Pr(Y^* = y^* \mid Y = y) \frac{\sum_{y' \in y^*} \Pr(Y = y')}{\sum_{y' \in y^*} \Pr(Y = y')} \\ &= \Pr(Y^* = y^* \mid Y = y)\end{aligned}$$

where terms with $\Pr(Y = y') = 0$ are zero and where the second equality uses (2.28). Note that the denominator must be positive since $\Pr(Y = y) > 0$ by assumption

Now (2.29) can equivalently be written as

$$\Pr(Y^* = y^* \mid Y = y \wedge Y \in y^*) = \Pr(Y^* = y^* \mid Y \in y^*)$$

since $\{Y = y\} \cap \{Y \in y^*\} = \{Y = y\}$ (and ignoring again y with $\Pr(Y = y) = 0$). In this form the condition can be re-expressed as a conditional independence condition:

$$\{Y = y\} \perp\!\!\!\perp \{Y^* = y^*\} \mid \{Y \in y^*\} \text{ or } \{Y^* = y^*\} \perp\!\!\!\perp \{Y = y\} \mid \{Y \in y^*\} \tag{2.30}$$

where $\perp\!\!\!\perp$ denotes stochastic independence and where the second version follows from the symmetry of conditional independence. Rewriting this again in the form of conditional probabilities one arrives at

$$\Pr(Y = y \mid Y^* = y^* \wedge Y \in y^*) = \Pr(Y = y \mid Y \in y^*)$$

But as a consequence of the consistency condition, if $\{Y^* = y^*\}$ obtains then so must $\{Y \in y^*\}$. Thus, $\{Y^* = y^*\} \subseteq \{Y \in y^*\}$ (see (2.3)) and so the previous equality is equivalent to

$$\Pr(Y = y \mid Y^* = y^*) = \Pr(Y = y \mid Y \in y^*)$$

which is just the first form of the CAR condition, (2.21). Therefore, all the conditions (2.21), (2.28), (2.29), and (2.30) are equivalent.

Gill et al. presented a further reformulation of CAR from which, as they informally argued, both the existence of CAR solutions without further conditions as well as the uniqueness on $\mathcal{D}(\mathcal{Y}^*)$ would follow (Gill et al. 1997: Sect. 2).

This reformulation starts from the consistency requirement $\{Y^* = y^*\} \subseteq \{Y \in y^*\}$ to arrive at

$$\begin{aligned}\Pr(Y^* = y^*) &= \Pr(Y^* = y^*, Y \in y^*)\\ &= \Pr(Y \in y^*)\Pr(Y^* = y^* \mid Y \in y^*)\end{aligned} \tag{2.31}$$

But the CAR version (2.29) requires that

$$\Pr(Y^* = y^* \mid Y \in y^*) = \Pr(Y^* = y^* \mid Y = y)$$

for all $y^* \in \mathcal{Y}^*$ and all $y \in y^*$ with $\Pr(Y = y) > 0$. It follows that one can write down the distribution of Y^* as a product of two terms

$$\Pr(Y^* = y^*) = P(y^*)\pi(y^*) \tag{2.32}$$

where $P(.)$ is a probability on $\mathcal{Y}$ and where the $\pi(y^*)$ satisfy

$$\sum_{y^*:y\in y^*} \pi(y^*) = 1 \quad \forall y \in \mathcal{Y}: \; \Pr(Y = y) > 0 \tag{2.33}$$

encoding version (2.29) of CAR.

Suppose that the decomposition (2.32) together with the restriction (2.33) holds. Then one may define the conditional distribution of Y given Y^* by

$$\Pr(Y = y \,|\, Y^* = y^*) := \frac{P(\{y\})}{P(y^*)}$$

for $y \in y^*$ and $\Pr(Y^* = y^*) > 0$. This is well defined since (2.32) together with $\Pr(Y^* = y^*) > 0$ implies $P(y^*) > 0$. If $y \notin y^*$, one sets $\Pr(Y = y \,|\, Y^* = y^*) = 0$.

Now the marginal distribution of Y is

$$\begin{aligned}
\Pr(Y = y) &= \sum_{y^*\in\mathcal{Y}^*} \Pr(Y^* = y^*)\Pr(Y = y \,|\, Y^* = y^*) \\
&= \sum_{y^*\in\mathcal{Y}^*} \Pr(Y^* = y^*)\frac{P(\{y\})}{P(y^*)} \\
&= \sum_{y^*\in\mathcal{Y}^*} P(y^*)\pi(y^*)\frac{P(\{y\})}{P(y^*)} \\
&= \sum_{y^*\in\mathcal{Y}^*} \pi(y^*)P(\{y\}) = P(\{y\})
\end{aligned}$$

since the sum over the $\pi(y^*) = 1$ by (2.33). Note that there is no need to define $\Pr(Y = y \,|\, Y^* = y^*)$ if $\Pr(Y^* = y^*) = 0$. It follows that

$$\begin{aligned}
\Pr(Y = y \,|\, Y^* = y^*) &= \frac{P(\{y\})}{P(y^*)} \\
&= \frac{\Pr(Y = y)}{\Pr(Y \in Y^*)} = \Pr(Y = y \,|\, Y \in y^*)
\end{aligned}$$

as long as $\Pr(Y^* = y^*) > 0$. If $\Pr(Y^* = y^*) = 0$, then either $P(y^*) = 0$ or $\pi(y^*) = 0$. In the first case, both sides can be left unspecified. If however $P(y^*) > 0$, the right hand side of the equality can be taken as a definition of the left hand side without affecting the joint distribution of (Y, Y^*). Lastly,

$$\begin{aligned}\Pr(Y \in Y^*) &= \sum_{y^* \in \mathcal{Y}^*} \Pr(Y \in y^* \mid Y^* = y^*) \Pr(Y^* = y^*) \\ &= \sum_{\substack{y^* \in \mathcal{Y}^* \\ \Pr(Y^* = y^*) > 0}} \frac{P(y^*)}{P(y^*)} \Pr(Y^* = y^*) = 1\end{aligned}$$

so that the consistency condition is met as well. In summary, if there is a decomposition of the distribution of Y^* as in (2.32) which obeys the restriction (2.33), then there is a joined distribution of (Y, Y^*) that is consistent and satisfies the CAR condition (2.21).

Conversely, suppose that the joined distribution of (Y, Y^*) satisfies the CAR condition (2.21) and that $\Pr(Y \in Y^*) = 1$. Then one can define

$$P(y^*) := \Pr(Y \in y^*) \quad \forall y^* \in \mathcal{Y}^*$$

and

$$\pi(y^*) := \begin{cases} \Pr(Y^* = y^* \mid Y \in y^*) & \Pr(Y^* = y^*) > 0 \\ 0 & \text{else} \end{cases}$$

With these definitions,

$$\Pr(Y^* = y^*) = P(y^*)\pi(y^*)$$

follows since

$$P(y^*)\pi(y^*) = \Pr(Y^* = y^* \wedge Y \in y^*) = \Pr(Y^* = y^*)$$

by consistency if $\Pr(Y^* = y^*) > 0$ or by definition if $\Pr(Y^* = y^*) = 0$. Moreover, for a fixed y with $\Pr(Y = y) > 0$

$$\sum_{y^* \ni y} \pi(y^*) = \sum_{\substack{y^* \ni y \\ \Pr(Y^* = y^*) > 0}} \Pr(Y^* = y^* \mid Y \in y^*)$$

$$= \sum_{\substack{y^* \ni y \\ \Pr(Y^*=y^*)>0}} \Pr(Y^* = y^* \mid Y = y)$$

$$= \sum_{y^* \ni y} \Pr(Y^* = y^* \mid Y = y) = 1$$

where the first equality uses the definition of $\pi(.)$, the second uses (2.29), the third the definition of conditional probability, and the last consistency again.

Thus (2.32) together with (2.33) are in fact equivalent to the other four forms of the CAR condition. The equations (2.32) are now linear in the probabilities of interest, $P(y^*)$ and the difficulties of solving the system are hidden in the restrictions (2.33) and the requirement that $P(.)$ must be a probability distribution.

There is a direct interpretation of the reformulated CAR conditions (2.32) and (2.33). Since $\pi(y^*)$ must equal $\Pr(Y^* = y^* \mid Y = y)$ for all $y \in y^*$ with $\Pr(Y = y) > 0$, the second term in the decomposition encodes the way incomplete data are constructed in the model, at least when $\Pr(Y = y) > 0$: Given $\{Y = y\}$, choose Y^* according to this conditional distribution. The decomposition thus states that the distribution of Y^* factors in a term that only depends on the distribution of interest (the distribution of Y), and in a term that describes the generation of the incomplete version Y^*. This may be translated into the statistical language by writing down the expected log-likelihood (with respect to the distribution of Y^*) for the parameters of interest $p_y := \Pr(Y = y)$ and the nuisance parameters π_{y^*}:[38]

$$\begin{aligned}
&\ell(p_y, \pi_{y^*}; y \in \mathcal{Y}, y^* \in \mathcal{Y}^*) \\
&= \sum_{y^* \in \mathcal{Y}^*} \Pr(Y^* = y^*) \log\big(P(y^*)\pi_{y^*}\big) \\
&= \sum_{y^* \in \mathcal{Y}^*} \Pr(Y^* = y^*) \log \sum_{y \in y^*} p_y + \sum_{y^* \in \mathcal{Y}^*} \Pr(Y^* = y^*) \log \pi_{y^*}
\end{aligned} \tag{2.34}$$

[38] I write here π_{y^*} instead of $\pi(y^*)$ to emphasise their role as a set of parameters. These π_{y^*} should not be confused with the ame symbol used in section 2.1.

where the summands are taken to be 0 if $\Pr(Y^* = y^*) = 0$. The likelihood factors into a part that depends on the parameters of interest only and in a part that depends on the nuisance parameters. The log-likelihood can be maximised separately in both parts subject to the restriction (2.33) and the requirement that p_y be a probability density. In consequence, if one is only interested in the density p_y one may simply maximise the first term of the log-likelihood ignoring the way incomplete data distributions are supposed to have emerged.[39]

This reformulation of the CAR conditions into statistical language provides an alternative though somewhat tedious way to show that to any given distribution of Y^* one can always construct a joint distribution of (Y, Y^*) such that both $\Pr(Y \in Y^*) = 1$ and CAR holds. It is nevertheless instructive to follow through the proof suggested by Gill et al. (1997) since as a consequence of their proof they claim uniqueness of the CAR solution for all sets y^* such that $\Pr(Y^* = y^*) > 0$. It is the uniqueness that was disputed at the end of the last section. And as I have indicated, non-uniqueness raises potentially serious problems for both the application of the CAR condition and the general acceptability of the "simple device".

Note first that if there is a decomposition of the distribution of Y^* as in (2.32) that satisfies the constraint (2.33) and such that $P(.)$ is a probability distribution, then this decomposition certainly maximises the expected likelihood (2.34). This is a simple consequence of Jensen's inequality[40] since for any other values p'_y, π'_{y^*} that satisfy the restrictions one has

$$\ell(p'_y, \pi'_{y^*}; y \in \mathcal{Y}, y^* \in \mathcal{Y}^*) - \ell(p_y, \pi_{y^*}; y \in \mathcal{Y}, y^* \in \mathcal{Y}^*)$$

[39] This is rather obviously true only if the nuisance parameters π_{y^*} are not related to the parameters of interest p_y. This is sometimes called 'parameter distinctness'. The concept of weak exogeneity in econometrics (Engle et al. 1983) and that of a cut (Barndorff-Nielsen 1978: 50; Barndorff-Nielsen/Cox 1994: 38) relate parameter distinctness to inferential procedure. While frequentist inference procedures are generally free to define their parameters arbitrarily (which many see as a threat to such 'inferences'), the mere existence of nuisance parameters will influence any frequentist inference procedure. I am not going to discuss these problems here, but Jaeger (2005a) provides some details in the case of coarsened data.

[40] The special case is known as the information inequality in coding and information theory.

$$
\begin{aligned}
&= \sum_{\substack{y^* \in \mathcal{Y}^* \\ \Pr(Y^*=y^*)>0}} \Pr(Y^* = y^*) \log \frac{P(y^*)' \pi'_{y^*}}{P(y^*) \pi_{y^*}} \\
&\leq \log \sum_{\substack{y^* \in \mathcal{Y}^* \\ \Pr(Y^*=y^*)>0}} \Pr(Y^* = y^*) \frac{P(y^*)' \pi'_{y^*}}{P(y^*) \pi_{y^*}} \\
&= \log \sum_{\substack{y^* \in \mathcal{Y}^* \\ \Pr(Y^*=y^*)>0}} \Pr(Y^* = y^*) \frac{P(y^*)' \pi'_{y^*}}{\Pr(Y^* = y^*)} \\
&= \log \sum_{\substack{y^* \in \mathcal{Y}^* \\ \Pr(Y^*=y^*)>0}} P(y^*)' \pi'_{y^*} \leq 0
\end{aligned}
$$

The last inequality follows from

$$
\begin{aligned}
\sum_{y^* \in \mathcal{Y}^*} P(y^*)' \pi'_{y^*} &= \sum_{y^* \in \mathcal{Y}^*} \sum_{y \in \mathcal{Y}} \mathbb{1}[y^*](y) p'_y \pi'_{y^*} \\
&= \sum_{y \in \mathcal{Y}} p'_y \sum_{y^* \in \mathcal{Y}^*} \mathbb{1}[y^*](y) \pi'_{y^*} = \sum_{y \in \mathcal{Y}} p'_y = 1
\end{aligned}
$$

where the restriction (2.33) is used for π'_{y^*} and p'_y is assumed to be a probability density. Note that the second equality requires (2.33) to hold only for all $y \in \mathcal{Y}$ with $p'_y > 0$.

In the other direction, the log-likelihood certainly has a maximiser since it is continuous and the set of parameters $(p_y, \pi_{y^*}; y \in \mathcal{Y}, y^* \in \mathcal{Y}^*)$ satisfying the restrictions is convex and compact.

Starting with the first summand of the log-likelihood that involves only p_y, one may show using a Lagrange multiplier ensuring $\sum_y p_y = 1$ that a maximiser of this summand must satisfy

$$
\sum_{y^* \ni y} \frac{\Pr(Y^* = y^*)}{P(y^*)} = 1 \text{ if } p_y > 0 \tag{2.35}
$$

Thus one can define

$$
\pi_{y^*} := \frac{\Pr(Y^* = y^*)}{P(y^*)}
$$

for all those y^* that contain at least one y with $p_y > 0$. For such sets y^*, certainly $\Pr(Y^* = y^*) = P(y^*)\pi_{y^*}$ and

$$\sum_{y^* \ni y} \pi_{y^*} = 1 \quad \forall y : p_y > 0$$

so that (2.33) is satisfied. Moreover, one can extend the definition of π_{y^*} to all elements of $\mathcal{Y}^*$ by setting it to 0 on sets y^* that only contain y with $p_y = 0$. For these sets, both $P(y^*) = 0$ and $\Pr(Y^* = y^*) = 0$ so that the decomposition holds for all elements of $\mathcal{Y}^*$. This set of parameters is a maximiser of the expected log-likelihood simply because it is a decomposition of the probabilities $\Pr(Y^* = y^*)$.

In summary, the argument shows that there is always a maximum of the expected log-likelihood and this must entail a decomposition (2.32). In other words, for any given probability distribution $\Pr(Y^* \in .)$ there is always a CAR solution, a result that was already obtained along a different route in section 2.2.4.

But what does this method tell about the uniqueness of CAR? Gill et al. (1997: 264) claim that their method provides uniqueness of $P(.)$ at least for those y^* with $\Pr(Y^* = y^*) > 0$. Moreover, they claim that the restriction (2.33) can be extended to hold for all elements $y \in \mathcal{Y}$, whether or not $p_y > 0$. Unfortunately, both claims are unfounded and it turns out that the problem is related to both claims simultaneously.

The examples of the previous section can be used to illustrate what goes wrong. They show that there may be several maxima of the log-likelihood function in the restricted variables p_y and π_{y^*}. Moreover, not all of them admit an extension of the restrictions on π_{y^*} to all of $\mathcal{Y}$.

Consider again the Monty Hall problem. As previously noted, there are two solutions given by

1) $\Pr(Y = 2) = 1$ and $\pi_{\{1,2\}} = \pi_{\{2,3\}} = 1/2$

2) $\Pr(Y = 1) = \Pr(Y = 3) = 1/2$ and $\pi_{\{1,2\}} = \pi_{\{2,3\}} = 1$

when $\Pr(Y^* = \{1,2\}) = \Pr(Y^* = \{2,3\}) = 1/2$. Both solutions do provide a decomposition, since in both cases $\Pr(Y^* = \{1,2\}) = \Pr(Y \in$

$\{1,2\})\pi_{\{1,2\}}$ and similarly for $\Pr(Y^* = \{2,3\})$. Thus both do satisfy the first part of the CAR condition (2.32). Also, they satisfy the restriction on π_{y^*} since in the first case, where only $y = 2$ must be considered, $\pi_{\{1,2\}} + \pi_{\{2,3\}} = 1$ and in the second case for both $y = 1$ and $y = 3$ only one summand with value 1 exists. Thus the solutions satisfy both (2.32) and (2.33). And obviously both solutions lead to the same value of the expected log-likelihood.

Nothing of this contradicts the arguments given in the proof above. But then Gill et al. (1997: 263–264) try to show that it is possible to construct the π_{y^*} so that their sum over all y^* that contain a given y, whether or not $p_y > 0$, must equal 1. They try to construct such a π_{y^*} using an extension of (2.35) to the case $p_y = 0$. Note that this results in a much stronger restriction than the one used in (2.33). Nevertheless, if the extension was possible, then there were $|\mathcal{Y}|$ valid constraints on the π_{y^*}. Thus the $2^{|\mathcal{Y}|} - 1$ equations $\Pr(Y^* = y^*) = P(y^*)\pi_{y^*}$ are exactly matched by the $|\mathcal{Y}|$ parameters p_y and the $2^{|\mathcal{Y}|} - 1 - |\mathcal{Y}|$ parameters π_{y^*}, making uniqueness a reasonable conjecture.

However, such an extension is impossible in both solutions to the Monty Hall problem. In the first solution, if $y = 1$, then $\pi_{\{1,2\}}$ is the only set to which it belongs, but its value is 1/2. Similarly for $y = 3$. Note that in this situation one might be tempted to define $\pi_{\{1\}} = \pi_{\{3\}} = 1/2$ so that the argument goes through nevertheless. But this strategy does not work for the second solution where for $y = 2$ we have $\pi_{\{1,2\}} + \pi_{\{2,3\}} = 2$. Since the π_{y^*} must be non-negative, one can not extend $\mathcal{Y}^*$ so that the sum condition can be met.

The only situation where the strategy to prove uniqueness will work is when $\Pr(Y^* = \{y\}) > 0$ for all $y \in \mathcal{Y}$. In that case, the original Gill et al. proof shows that there must be a unique CAR distribution. It is still an open question how to characterise those situations where the only form of non-uniqueness of CAR is of the simple form Gill et al. conjectured. There are not even easily stated conditions which imply the existence of multiple solutions of the type exemplified by the Monty Hall problem and Example 3a.

2.2.7. Strong CAR?

With the present CAR condition and its various re-expressions there is always at least one joint distribution of (Y, Y^*) such that a given incompleteness $\{Y^* = y^*\}$ implies nothing more about the distribution of interest than that $\{Y \in y^*\}$ (this is condition (2.21)) or such that incompleteness does not depend on the actual value of the random variable Y within the coarsening set y^* (condition (2.28)), or that the coarsened variables are constructed without regard to particular values of the variable of interest (condition (2.30)), or that the likelihood factors into a part describing the distribution of interest and the 'missing mechanism' (condition (2.32)). All these CAR formulations turned out to be equivalent mathematical conditions for which a solution always exists. That is, for any given distribution of Y^* there is a joined distribution of (Y, Y^*) that is consistent and compatible with the CAR condition. The CAR condition does not restrict the set of possible distributions of the coarsened variable.[41]

This is what one ought to hope for since the "simple device" can prove useful only to the extend that its central concepts are not special mathematical artifacts. The definition of CAR depends crucially on the introduction of a probability model. Its usefulness as a regulative idea would be impaired if its definition would depend on particular data constellations.

On the other hand, the CAR condition may lead to contradictory results even for all non-trivial $y^* \in \mathcal{Y}^*$, as Example 3a and the Monty Hall problem demonstrate. This situation is at least as threatening to the use of CAR since it would imply contradictory results for many or even all sets of interest. One possible solution, albeit a rather radical one, is to strengthen the CAR definition in such a way that contradictory

[41] This is no longer true when infinite sets are considered. See Gill et al. (1997: 264) for an example. They conjecture that this problem can be alleviated by "compactifying both the sample space and all the observed random sets in a careful way". A discussion of this type of difficulty will not be pursued here. Cator (2004: 1962) provides another type of example where the difficulty lies in the fact that the incompleteness cannot be represented in terms of subsets of the set $\mathcal{Y}$. A discussion of this type of problem will be postponed to the next section.

results are ruled out. One may then hope that the situations that are not CAR according to the strengthened definition in fact turn out to be 'pathological' exceptions. An attractive possibility is opened up by the CAR formulations of the previous subsection.

While all these definitions are equivalent with CAR in the form (2.21), the change of the conditioning variable from Y^* in (2.21) to Y suggests that one may ask for (2.28)–(2.30) to hold for all y and not only for those with $\Pr(Y = y) > 0$. This is a vacuous strengthening if $\Pr(Y^* = \{y\}) > 0$ for all $y \in \mathcal{Y}$ because then $\Pr(Y = y) > 0$ by the consistency requirement. In all other cases, however, it is strictly stronger than the CAR condition (2.21). This is an attractive approach since one may argue that (2.28)–(2.30) refer to a 'mechanism' producing the form of incompleteness from any given 'input' y. If such a 'mechanism' can reasonably be posited, it should not depend on whether or not in fact $\Pr(Y = y) > 0$ holds true. It should be possible to define the 'mechanism' of missingness without paying attention to particulars of a given situation. They ought to be irrelevant for the definition of a predetermined procedure.

The exclusion of situations where such a 'mechanism' does not exist and thus the argument would have no force might be tolerable if this would happen only in rare or exceptional circumstances. But one of the arguments of the critics of the "simple device" must be born in mind. It challenges that there is such a thing as a probabilistic 'mechanism' or an 'answer probability' of interviewees in a survey at all. Such a position has some empirical support for social science surveys, and it certainly pertains to observational studies where the degree of incomplete information depends on the deliberation of respondents.

On the other hand, the modelling strategy might have some value, at least if it does not presuppose facts about surveys or the interviewees. It should only exclude unreasonable mathematical artifacts from consideration that would not have been considered by informed survey statisticians either. It would then not touch directly on the question whether the 'mechanism' must be supposed to 'work' for every interviewee, or whether its preconditions are ever encountered by any member of the population.

In short, the suggestion is to require that (2.28) and the other versions should hold not only for all y that actually occur but for all $y \in \mathcal{Y}$. This version of CAR has been termed *strong* CAR by Jaeger (2005a) and Grünwald/Halpern (2003). If this direction is pursued, then (2.28) is strengthened to

$$\Pr(Y^* = y^* \mid Y = y) \text{ is constant on } y \in y^* \tag{2.36}$$

for all $y^* \in \mathcal{Y}^*$ and all $y \in y^*$, even when $\Pr(Y = y) = 0$.

It turns out that both the Monty Hall example and Example 3a do not satisfy strong CAR. This follows in the case of the Monty Hall example along the same lines as indicated at the end of section 2.1.2: If the price is behind door 1 (and the contestant has chosen door 2), then Monty is forced to open door 3. Thus $\Pr(Y^* = \{1,2\} \mid Y = 1) = 1$. From strong CAR, it follows that $\Pr(Y^* = \{1,2\} \mid Y = 2) = 1$. But $\Pr(Y^* = \{2,3\} \mid Y = 3) = 1$ as well, from which $\Pr(Y^* = \{2,3\} \mid Y = 2) = 1$ would follow. In consequence, $1 = \Pr(Y^* \in \mathcal{Y}^* \mid Y = 2) = \Pr(Y^* = \{1,2\} \mid Y = 2) + \Pr(Y^* = \{2,3\} \mid Y = 2) = 2$, a contradiction. The contradiction is only avoided when $\Pr(Y = 2) \in \{0,1\}$ *and* simultaneously conditional distributions with conditioning events whose probability is zero are not assumed to take a definite value.

Similarly, in Example 3a

$$\begin{aligned} \Pr(Y^* = \{2,3\} \mid Y = 2) &=: a = \Pr(Y^* = \{2,3\} \mid Y = 3) \\ \Pr(Y^* = \{1,2,3\} \mid Y = 2) &=: b = \Pr(Y^* = \{1,2,3\} \mid Y = 3) \end{aligned}$$

so that $a + b = 1$. But this forces $\Pr(Y^* = \{1,2\} \mid Y = 2) = 0 = \Pr(Y^* = \{1,2\} \mid Y = 1)$, contradicting $\Pr(Y^* = \{1,2\}) = 0.1 > 0$.

In general, strong CAR will be impossible whenever there are properly nested edges in the hypergraph of $(\mathcal{Y}, \mathcal{Y}^*)$. In that case, the sum over all nodes of the conditional distributions given the inner edge must equal the sum of the conditional distributions given the outer edge and both must equal 1. But this can not be the case when the nodes unique to the outer edge have non-vanishing probability. For suppose that all elements of $\mathcal{Y}^*$ have positive probability and that for some $y, y' \in \mathcal{Y}$

$$A := \{y^* \in \mathcal{Y}^* \mid y \in y^*\} \subsetneq \{y^* \in \mathcal{Y}^* \mid y' \in y^*\} =: B$$

so that the edge of the hypergraph corresponding to y is nested within the one corresponding to y'. Then $\Pr(Y^* = y^* \mid Y = y) = \Pr(Y^* = y^* \mid Y = y')$ and $\sum_{y^* \in A} \Pr(Y^* = y^* \mid Y = y) = 1$. It follows that $\Pr(Y^* = y^* \mid Y = y') = 0$ for all $y^* \in B \setminus A$. Fix one such y^*, say $y^{*\prime}$. By the strong version of (2.29) $\Pr(Y^* = y^{*\prime} \mid Y \in y^{*\prime}) = 0$. But now $\Pr(Y^* = y^{*\prime}) = 0$ follows from consistency, in contradiction to the assumption that all elements of $\mathcal{Y}^*$ have positive probability. Thus the existence of properly nested edges of the hypergraph implies that strong CAR does not hold.

In the Monty Hall example, strong CAR fails and in that setting, the idea of a well defined 'mechanism' that produces the coarsened data can be based on the expectation that the rules of the game are in fact well defined for all situations that might arise. In this case the argument for strong CAR based upon a 'mechanism' defined independently of the distribution of Y has much force. Similarly, in Example 3a, the contradictory solutions $(5/6, 1/6, 0)$ and $(6/7, 0, 1/7)$ should be ruled out as well. But strong CAR excludes many more situations. Consider e.g. Example 3. Here there is a unique CAR solution $(5/7, 2/7, 0)$, but the hypergraph given in Figure 2.3 shows nested edges for $Y = 2$ and $Y = 3$ so that there is no strong CAR solution. On the other hand, the interpretation given at the beginning of section 2.1 can certainly not be ruled out by referring to a 'mechanism'. It would be simply too much to ask for a 'mechanism' that produces coarsened data according to strong CAR in a survey context. Even if the illusion of respondents answering according to some probability model would be taken for granted, it would be unreasonable to expect that they should be able to react to circumstances and situations that nobody ever encountered. But this would be needed to define a 'mechanism'.

There is another argument in favour of strong CAR. It is based on the fact that a CAR solution that is not strong CAR as well must be such that $\Pr(Y \in D) = 0$ for some element $D \in \mathcal{D}(\mathcal{Y}^*)$. As Examples 1a and 5 demonstrate, strong CAR may hold even when some solutions imply 0 probabilities for a subset of $\mathcal{Y}$. Moreover, $\Pr(Y^* = y^*) > 0$ for all $y^* \in \mathcal{Y}^*$ implies $\Pr(Y \in y^*) > 0$ by consistency (2.6). The difference between CAR and strong CAR arises solely if CAR implies 0 probability for at least one element $D \in \mathcal{D}(\mathcal{Y}^*) \setminus \mathcal{Y}^*$.

Grünwald and Halpern (2003) have argued that degenerate solutions (zero probabilities for some sets) are the main source of difficulties encountered with CAR solutions. They counter the slogan 'Car is everything' by 'sometimes CAR is nothing' (Grünwald/Halpern 2003: 256).[42] There are obviously circumstances where degenerate results indicate doubtful results. If to a given $\mathcal{Y}$ a further element, say $*$ is added and if $\mathcal{Y}^*$ is modified so that all y^* contain $*$, then obviously $*$ can not be distinguished from any other element of $\mathcal{Y}$.[43] But the CAR condition now leads to either $\Pr(Y = *) = 1$ or $\Pr(Y = *) = 0$. Moreover, there is no strong CAR solution since all other edges of the hypergraph are nested within that of $*$, and some will be properly nested. Here CAR would seemingly state much more than just that conditioning on $\{Y^* = y^*\}$ can be replaced by conditioning on $\{Y \in y^*\}$. It says something about the existence (or non-existence) of certain events. From a Bayesian perspective, one is forced to assign probability zero to events which were a priori considered possible. The disturbing feature of the example is that from (weak) CAR there seems to follow something definite even though nothing must be known or assumed about the particular situation, not even probabilities for the coarsened variables. This seems to suggest that there is more in (weak) CAR than the assertion that coarsened variables can be taken at face value.

But adding an extra symbol $*$ to a model may be ruled out either by an appeal to parsimony or because it would generally lead to ridiculous models. There is no need to apply to strong CAR. Consider the survey statistician who, after the survey, considers that labour market conditions of many respondents are much more complicated than the simple distinction between 'employment', 'out of labour force', and 'unemployment'. He could add further categories to his $\mathcal{Y}$ and surmise that respondents are kind enough to answer to the original question even if they do not consider themselves to be in any of the categories that were being asked. Redoing his analysis he finds that either nobody or everybody was in his additional categories just by invoking CAR. Since this can be done with

[42] Grünwald and Halpern (2003: Ex. 4.6) provide examples where the degenerate solutions do not arise from incomplete identification and characterise such situations for the case $|\mathcal{Y}^*| = 3$.

[43] This is a form of Jaeger's (2005a) second example.

all categories one might ever come up with, the result of such a strategy is self-defeating.

The argument concerning degenerate solutions is compelling in the case of an additional element $*$ or in the Monty Hall example. But in cases like Example 3 it has no more force than the argument based on 'mechanisms'. Note that in Example 3, CAR always forces $\Pr(Y = 3) = 0$ as long as $\Pr(Y^* = \{3\}) = 0$ (the structure of $\mathcal{Y}^*$ does not change) and the other probabilities of elements in $\mathcal{Y}^*$ are positive. And $\{3\} = \mathcal{Y} \setminus \{1, 2\} \in \mathcal{D}(\mathcal{Y}^*) \setminus \mathcal{Y}^*$ so that CAR puts 0 probability on the 'wrong' subset. But that CAR implies a particular distribution only shows that CAR is a very strong condition: It can identify a single distribution from a large set of consistent distributions. That this particular solution puts 0 probability on some elements of $\mathcal{Y}$ is as such no argument against that solution.

In summary, the CAR condition in the form (2.21) can be satisfied for any given distribution on $\mathcal{Y}^*$. Therefore, weak CAR is untestable as it should be if the modelling strategy is not to be an artifact of the "simple device". On the other hand, strong CAR may fail. Note that the strong version of CAR in (2.36) refers to conditional distributions and is required to hold for all $y^* \in \mathcal{Y}^*$ and all $y \in y^*$. The condition must hold irrespective of whether $\{Y = y\}$ has positive probability or not. But for those points $y \in \mathcal{Y}$ that have 0 probability, the joint probability of $\{Y = y, Y^* = y^*\}$ must be 0 as well. But then the conditional probabilities in the condition (2.36) are not needed to define the joined distribution. Using (2.36) or equivalent formulations as a CAR condition thus uses more than just the joined distribution of (Y, Y^*).

One possible justification of such a move is the requirement of a 'mechanism' of coarsening whose definition should not depend on any aspects of Y. Another argument is based on the possibly strong ontological commitment that may follow from degenerate probability assignments. But in many relevant cases in the social sciences, neither argument is compelling. In these cases, strong CAR seems to be much too strong, excluding even situations as Example 3 which otherwise looks perfectly natural. Moreover, strong CAR is not neutral as a regulative modelling strategy since the structure of $(\mathcal{Y}, \mathcal{Y}^*)$ on which it hinges is at least partly determined by the data.

Lastly, the main problem with weak CAR is not the existence of degenerate solutions for elements of $\mathcal{D}(\mathcal{Y}^*)$, the situation addressed by both arguments. It is the existence of contradictory solutions as in the case of Example 3a and the Monty Hall problem where there are two values of $\Pr(Y \in D)$ for some $D \in \mathcal{D}(\mathcal{Y}^*)$. This follows from the existence of disconnected solution sets implied by weak CAR. Excluding such situations, a condition strictly weaker than the strong CAR condition, may turn out to be a reasonable compromise that almost saves the slogan 'CAR is everything'.

2.3. Coarsening Variables and Coarsening Completely at Random

Even if the probability model is taken as the natural framework to discuss incomplete observations in a general context, it is by no means implied that the formalisations of the concept of coarsening at random, and therefore an appreciation of departures from it, are to be framed in terms of conditional distributions as was done up to now. In the literature on missing data problems one finds many variants of the missing at random conditions, often formulated after introducing additional random variables. Before the CAR conditions can justifiably be used within the probability model, the status of different versions of the conditions must be discussed. In particular, it must be shown that no further structure is implied by different formulations of the conditions.

Of particular importance is the relation of the present formulation to the now classical formulation used by Little and Rubin (2002). The latter exploits additional random variables that together with the variable of interest determine the incomplete variable Y^*. Then the missing at random (or coarsening at random) conditions are expressed in terms of conditional densities involving the additional variables.[44]

[44] It seems that the origins of this way to formulate missing or coarsening at random conditions must go back at least to the early 20th century. It is used implicitly in Fisher's 1925 article on Maximum Likelihood, inter alia.

The next subsection discusses the relation of this classical formulation to the formulation used in the previous sections. It turns out that the classical formulation potentially adds further structure to the "simple device" so that the two formulations may provide contradictory results. Furthermore, the classical formulation allows to define the concept of "completely coarsening at random" or CCAR rather naturally. But it turns out that the concept is rather difficult to formulate in general. Some of the consequences are discussed in section 2.3.2. On the other hand, the additional structure can sometimes be exploited to express information present in incomplete data problems that can not be expressed in the simple framework used previously. Moreover, there are situations that are naturally described as incomplete data problems but that can not be expressed in the language of subsets of the original sample space. This is discussed in the last subsection.

2.3.1. Coarsening Variables

The early development of methods for censored duration data was based nearly exclusively on the random censorship model where a censoring variable, say R, was stipulated such that the coarsened (censored) observations are represented by $(Y', \delta) := (\min(Y, R), \mathbb{1}[Y < R])$. There is a natural one-to-one translation into the previous setup by setting $y^* = \{y'\}$ if $\delta = 1$ and $y^* = \{y', \ldots\}$ otherwise. In the other direction, one would put $y' = \min y^*$ and $\delta = 0$ if $|y^*| > 1$, and $y' = \min y^*$, $\delta = 1$ otherwise. But it is not so obvious what the role of R is. In particular, its value is not determined by the value of y^*: If $(Y' = y', \delta = 1)$, the value of R can not be inferred. Its inclusion extends the setup previously discussed and it is plain that the additional structure needs justification.

In the case of missing data, the additional variable is an indicator, say R, which takes the value 1 if the data are complete, 0 otherwise. The incomplete variable is then a function $Y^* = g(Y, R)$ such that $g(y, 1) = \{y\}$ and $g(y, 0) = \mathcal{Y}$. Knowing the underlying value of Y and the realised value of R, the value of the observation can be computed by a known function. In this case, one can indeed infer the value of r from the value of y^* so that nothing essential is added to the structure.

Rubin’s article of 1976, following the tradition, uses the existence of an indicator R to define his version of MAR. In general, the idea is to express MAR/CAR as a property of a procedure that from the random variable Y plus a randomisation variable R produces the observed Y^*. Such a formulation is obviously very closely connected to the idea of a ‘mechanism’ and the formulation of CAR in (2.28).

The method can easily though rather vacuously be adapted to deal with all partially complete data defined as set-valued observations. To see this, define a random variable R as

$$R\colon \Omega \longrightarrow \mathcal{R} \coloneqq \mathcal{Y}^* \tag{2.37}$$

where for ease of translation the range of R is chosen to equal $\mathcal{Y}^*$. Furthermore, the conditional distribution of R given Y can be defined as

$$\Pr(R = y^* \mid Y = y) \coloneqq \Pr(Y^* = y^* \mid Y = y) \tag{2.38}$$

for all y such that $\Pr(Y = y) > 0$. This defines a joined distribution of (Y, R). Now the function $g(.,.)$ can be defined as

$$\begin{aligned} &g\colon \mathcal{Y} \times \mathcal{R} \longrightarrow \mathcal{Y}^* \\ &g(y, r) \coloneqq r \end{aligned} \tag{2.39}$$

using just the second argument. Since (Y, Y^*) is consistent, $\Pr(Y \in g(Y, R)) = \Pr(Y \in R) = \Pr(Y \in Y^*) = 1$ so that (Y, R, g) is consistent as well.

This is a trivial reformulation of the setup in the previous sections, even though it introduces further elements that seem to be extraneous to the situations. However, it is probably the easiest way to express the idea that the incomplete data arise from the underlying data by some randomisation expressed by R. Moreover, it may be used as a first translation of the previous formulation into the more familiar formulation used in the classical literature. The construction shows that to each incomplete data model in terms of (Y, Y^*) there is an equivalent one in terms of (Y, R, g) though the translation need not be unique.

In the other direction, to any triple (Y, R, g) there is a tuple (Y, Y^*) defined by $(Y, Y^*) \coloneqq (Y, g(Y, R))$. One may call the joined distribution of (Y, R) CAR if the joined distribution of (Y, Y^*) implied by (Y, R) and $g(.,.)$ is CAR. It is now natural to inquire whether CAR can be directly formulated in terms of (Y, R, g). Using (2.28) as a natural starting point, (Y, R, g) might tentatively be called CAR if for all $r \in \mathcal{R}$ and all y of positive probability

$$\Pr(R = r \mid Y = y') \text{ is constant on } \{y' \mid y' \in g(y, r), \Pr(Y = y') > 0\} \tag{2.40}$$

With the trivial construction of a (Y, R, g) from the tupel (Y, Y^*) given above, (2.40) certainly implies CAR. However, the appeal of the classical formulation derives from a much simpler structure of R. In the case of missing data, R is just an indicator. In the case of censored data, it is a number giving censoring times. Both versions of R are much easier to handle than the brute force translation that simply identifies R with the set-valued variable Y^*. Moreover, the classical formulation justifies the use of the face value likelihood by invoking stochastic independence of their R from the underlying Y, a condition that is much easier to treat than the respective CAR conditions or the clumsy (2.40).

However, this ease of characterising CAR can be deceptive in general incomplete data models. Returning again to Example 1, one might define a random variable R taking the values 0, 1, 2 that specifies the degree of incompleteness. Then one can define the variable $Y^* := g(Y, R)$ via

$$\begin{aligned} g(y,2) &= \{y\} \\ g(y,1) &= \begin{cases} \{2,3\} & y \in \{2,3\} \\ \{1,2,3\} & y = 1 \end{cases} \\ g(y,0) &= \mathcal{Y} \end{aligned} \tag{2.41}$$

With this definition of $g(.,.)$ and $\mathcal{R}$ one recovers the joined CAR distribution of (Y, Y^*) by setting the joined distribution of (Y, R) to

	Y	1	2	3	
R	0	5/80	3/80	3/80	11/80
	1	5/80	4/80	4/80	13/80
	2	1/2	8/80	8/80	56/80
		10/16	3/16	3/16	

Note that the joined distribution of (Y, R) and therefore the marginal of R is not unique for one might as well choose

	Y	1	2	3	
R	0	0	3/80	3/80	6/80
	1	1/8	4/80	4/80	18/80
	2	1/2	8/80	8/80	56/80
		10/16	3/16	3/16	

without changing the joint distribution of (Y, Y^*).

Obviously, the choice of $g(.,.)$ is not unique either, even when the setting of $R \coloneqq Y^*$ is discarded in favour of more concise representations. In Example 1, another possible choice is

$$\begin{aligned} g(y,2) &= \{y\} \\ g(y,1) &= \begin{cases} \{2,3\} & y \in \{2,3\} \\ \{1\} & y = 1 \end{cases} \\ g(y,0) &= \mathcal{Y} \end{aligned} \tag{2.42}$$

where now the joined distribution of (Y, R) might be

	Y	1	2	3	
R	0	1/8	3/80	3/80	16/80
	1	0	4/80	4/80	8/80
	2	1/2	8/80	8/80	56/80
		10/16	3/16	3/16	

The obvious non-uniqueness of the (Y, R, g) representation of incomplete data creates a problem for the definition of CAR in terms of a missingness indicator. If the first choice of $g(.,.)$ is combined with either of the two joined distributions of (Y, R) proposed above, then both violate the condition (2.40), because

$$\Pr(R = 0 \mid Y = 1) = \frac{5/80}{50/80} = \frac{1}{10} \neq \Pr(R = 0 \mid Y = 2) = \frac{3/80}{15/80} = \frac{1}{5}$$

in the first case and

$$\Pr(R = 0 \mid Y = 1) = 0 \neq \Pr(R = 0 \mid Y = 2) = \frac{3/80}{15/80} = \frac{1}{5}$$

in the second case. Thus there are (Y, R, g) that imply a CAR distribution for (Y, Y^*) but that are not CAR according to (2.40).

On the other hand, the second choice of $g(.,.)$ yields

$$\Pr(R = 0 \mid Y = 1) = \frac{2/16}{10/16} = \frac{1}{5} = \Pr(R = 0 \mid Y = 2) = \frac{3/80}{3/16}$$

and similarly for $\Pr(R = 0 \mid Y = 3)$. Furthermore,

$$\Pr(R = 1 \mid Y = 2) = \frac{4/80}{3/16} = \frac{4}{15} = \Pr(R = 1 \mid Y = 3)$$

covering all cases required by (2.40). Not too surprisingly, therefore, condition (2.40) strongly depends on the choice of R and $g(.,.)$. Even if the joined distribution of (Y, Y^*) is CAR according to the definitions in the previous sections, this need not be the case according to 2.40 based on the conditional distribution of R given Y.

The case of missing data is an exception in that in this simple case the formulations in terms of a missing indicator and in terms of the tuple (Y, Y^*) are in fact equivalent. When Heitjan and Rubin (1991) introduced the term CAR, they defined it in terms of the joined distribution of (Y, Y^*) using (2.28) as definition, even though the coarsening was defined by a function such that $Y^* = g(Y, R)$. Their corollary 1 then stated that for 'consistent' $g(.,.)$ and those R that are functions of Y^* the formulation (2.40) is equivalent to CAR.

There have been several attempts to explicate 'consistent' choices of $g(.,.)$ such that the relation to CAR can be stated in terms of (Y, R, g). First, the consistency requirement $\Pr(Y \in Y^*) = 1$ translates into

$$\Pr(Y \in g(Y, R)) = 1 \tag{2.43}$$

If $\mathcal{Y}$ is finite, then this is equivalent to

$$\Pr(g(y, R) \ni y \mid Y = y) = 1 \tag{2.44}$$

for all $y \in \mathcal{Y}$ with $\Pr(Y = y) > 0$. If the roles of Y and R are reversed, one obtains

$$\Pr(Y \in g(Y, r) \mid R = r) = 1 \tag{2.45}$$

for all $r \in \mathcal{R}$ such that $\Pr(R = r) > 0$.

Jaeger (2005b: Definition 2.5) explicitly strengthened the condition to

$$y \in g(y, r) \tag{2.46}$$

for all y such that $\Pr(Y = y) > 0$ and all $r \in \mathcal{R}$. This is a strictly stronger condition than (2.43). In particular, (2.46) is violated by the construction in (2.37)–(2.39). But it is sometimes easier to work with this deterministic condition than with (2.43). Therefore much of the literature concentrates on this case.[45]

Jaeger (2005b: Definition 2.5) further assumes that

$$y' \in g(y, r) \Longrightarrow g(y', r) = g(y, r) \tag{2.47}$$

for all r and all y, y' of positive probability. Note that (2.47) holds trivially for the construction (2.37)–(2.39) since $g(.,.)$ depends only on r. Both conditions also hold for the case of missing and censored data when the latter are expressed in terms of y^* instead of the more usual (y', δ).

However, the conditions severely restrict the models that can be represented by the triple (Y, R, g). In fact, the set $\mathcal{Y}^*$ corresponding to such an (Y, R, g) must be a union of partitions of $\mathcal{Y}$. To see this, fix an $r \in \mathcal{R}$ and set

$$\mathcal{Y}^*_r \coloneqq \{y^* \in \mathcal{Y}^* \mid \exists y \in \mathcal{Y} : g(y, r) = y^*\}$$

But $g(., r)$ trivially induces a partition on $\mathcal{Y}$ and the conditions allow to identify it with $\mathcal{Y}^*_r$. More precisely, by condition (2.46)

$$\mathcal{Y} = \bigcup_{y \in \mathcal{Y}} g(y, r) = \bigcup_{y^* \in \mathcal{Y}^*_r} y^*$$

[45] Heitjan and Rubin (1991: 2247) use a similar condition. See Gill et al. (1997: 282) for some remarks.

Suppose further that there is a $y \in y_1^* \cap y_2^*$ for $y_1^*, y_2^* \in \mathcal{Y}_r^*$. Then $g(y,r) = y_1^*$ since there is a y' with $g(y',r) = y_1^*$ by the definition of $\mathcal{Y}_r^*$ and by (2.47), $g(y,r) = g(y',r) = y_1^*$. Similarly, $g(y,r) = y_2^*$. Consequently, $y_1^* = y_2^*$. Therefore, $\mathcal{Y}_r^*$ is a partition of $\mathcal{Y}$ and hence $\mathcal{Y}^* = \cup_r \mathcal{Y}_r^*$ a union of partitions. Thus, with the restrictions (2.46) and (2.47), the variable R simply indicates a particular partition chosen to produce the coarsening y^*.

Even though the restrictions (2.46) and (2.47) exclude $\mathcal{Y}^*$ that are not unions of partitions of $\mathcal{Y}$ (e.g. Example 3), they provide a closer connection between the (Y, R, g) formulation and the implied pair (Y, Y^*). So one might try to define the CAR condition in terms of (Y, R, g) by

$$\Pr(R = r \mid Y = y) \text{ is constant on} \tag{2.48}$$

$$\{y \mid g(y,r) = y^*, \Pr(Y = y) > 0\} \tag{2.49}$$

or

$$\Pr(R = r \mid Y = y) = \Pr(R = r \mid Y \in \{y \mid g(y,r) = y^*\}) \tag{2.50}$$

or

$$\{Y = y\} \perp\!\!\!\perp \{R = r\} \mid \{Y \in \{y \mid g(y,r) = y^*\}\} \tag{2.51}$$

for all $y^* \in \mathcal{Y}^*$ where the conditions are only required to hold for y of positive probability. These conditions parallel the CAR conditions formulated in (2.28), (2.29), and a combination of (2.21) and (2.29), respectively. I will call any of these equivalent conditions R-CAR.

The example of $g(.,.)$ in (2.41) does not satisfy condition (2.47), but (2.42) satisfies both conditions (2.46) and (2.47). It also satisfies the R-CAR condition (2.48) as shown above. Thus one might hope that the implied tuple (Y, Y^*) is CAR if there exists an (Y, R, g) model satisfying (2.46) and (2.47) that is R-CAR. This is in fact true, since

$$\begin{aligned}\Pr(Y^* = y^* \mid Y = y) &= \Pr(R \in \{r \in \mathcal{R} \mid g(y,r) = y^*\} \mid Y = y)\\ &= \sum_{r : g(y,r) = y^*} \Pr(R = r \mid Y = y) \end{aligned} \tag{2.52}$$

But by (2.47), $g(y', r) = y^*$ for all $y' \in y^*$ so that the summation is over the same set of values of r. From R-CAR (2.48), the $\Pr(R = r \mid Y = y)$ are constant on y^*. It follows that $\Pr(Y^* = y^* \mid Y = y)$ is constant on $y \in y^*$ and thus CAR.

Jaeger (2005b: Theorem 2.9) claimed that the converse holds as well, namely that if the distribution of (Y, Y^*) is CAR, then there is an R-CAR model. But this is wrong since the conditions (2.46) and (2.47) rule out the representation of CAR models where $\mathcal{Y}^*$ is not a union of partitions. As Example 3 demonstrates, even in this case there is a joined distribution of (Y, Y^*) that is CAR. The inconsistency arises because Jaeger uses the trivial (Y, R, g) construction given at the beginning of this section to prove the existence of R-CAR. However, this construction violates restriction (2.46) which Jaeger presupposes in his Definition 2.5.

There is another way in which (Y, R, g) representations together with the conditions (2.48)–(2.51) may fail to be equivalent to CAR. The restrictions (2.46) and (2.47) insure that $g(.,.)$ is sensitive to the y argument so that examples like the trivial construction at the beginning of the section are ruled out. In that case, $g(.,.)$ did not even depend on y. But it may as well be that $g(.,.)$ is insensitive to values of r. In that case, the conditions (2.48)–(2.51) would be either void (using any independent distribution of R) or false (using a distribution of R contradicting (2.48)–(2.51)) without changing the implied joined distribution of (Y, Y^*). A simple example was provided by Jaeger (2005b: 895). Suppose that $Y = (Y_1, Y_2)$ is a two-dimensional binary variable and that the corresponding $\mathcal{Y}^*$ is just $\{\{(0,0), (0,1)\}, \{(1,0)\}, \{(1,1)\}\}$. Here, Y_2 is missing if $Y_1 = 0$. Otherwise, all values are present. Now one may choose

$$\begin{aligned} g((0,0), r) &\coloneqq g((0,1), r) \coloneqq \{(0,0), (0,1)\} \\ g((1,0), r) &\coloneqq \{(1,0)\} \\ g((1,1), r) &\coloneqq \{(1,1)\} \end{aligned}$$

so that $g(.,.)$ does not depend on r at all. If $\mathcal{R} = \{r\}$, then (2.48)–(2.51) are trivially true. But with $\mathcal{R} = \{r_1, r_2\}$ one may easily produce examples that falsify the R-CAR conditions. E.g., one might choose

$$\Pr(R = r_1 \mid Y = (1,0)) = \Pr(R = r_1 \mid Y = (1,1)) = 1$$

$$\Pr(R = r_1 \mid Y = (0,1)) = 1/3, \quad \Pr(R = r_1 \mid Y = (0,0)) = 2/3$$
$$\Pr(R = r_2 \mid Y = (0,1)) = 2/3, \quad \Pr(R = r_2 \mid Y = (0,0)) = 1/3$$

Since $g(.,.)$ does not depend on r, the implied distribution of (Y, Y^*) stays the same, whatever distribution is assumed for Y. Now

$$\Pr(R = r_1 \mid Y = (0,1)) = 1/3 \neq \Pr(R = r_1 \mid Y = (0,0)) = 2/3$$

and also

$$\Pr(R = r_2 \mid Y = (0,1)) = 2/3 \neq \Pr(R = r_1 \mid Y = (0,0)) = 1/3$$

so that (2.48) is violated.

In consequence, one might try to restrict the setup even further by asking that $g(y,.)$ is injective for all fixed y. Then r is a function of y^* such that for $y \in y^*$

$$y^* = g(y,r) \Longleftrightarrow r = h(y^*) \tag{2.53}$$

for some function $h(.)$.[46] In this case, (2.52) reduces to

$$\Pr(Y^* = y^* \mid Y = y) = \Pr(R = h(y^*) \mid Y = y)$$

so that if the left hand side is constant on y^* then so is the right hand side and vice versa. In this special case, CAR and R-CAR are equivalent: If there is a representation of a (Y, Y^*) pair in terms of a triple (Y, R, g) satisfying (2.46) and (2.47), and such that R is a function of Y^*, then the joined distribution of (Y, Y^*) is CAR if and only if (Y, R, g) is R-CAR.

However, this is a very special situation. Not only is $\mathcal{Y}^*$ the union of partitions indexed by r, say $\mathcal{Y}^*_r$, but the intersection of the $\mathcal{Y}^*_r$ must be empty. Measurements at different levels of precision, i.e. unions of strict refinements of partitions provide an example. On the other hand,

[46] If the condition holds, then Jaeger (2005b: 894) calls R 'observable'. But this is an unfortunate choice of term since it conflates the "simple device" and its modelling methods with the observations obtained in a survey.

$\mathcal{Y}^* = \{\{1,2,3\}\}, \{\{1\},\{2,3\}\}, \{\{1\},\{2\},\{3\}\}$ from Example 1 neither is a union of strict refinements nor does it permit a function $h(.)$ since $\{1\}$ appears in two of the partitions and there is no other way to arrange the partitions. Hence, a function $h(.)$ that would select a partition based only on y^* can not exist. Similarly, neither Example 2 nor Example 3 are of this form, nor are censored data.[47]

Thus, the CAR and R-CAR conditions are guaranteed to be equivalent only if R is a function of Y^*. But this condition severely restricts the structure of $\mathcal{Y}^*$. If the restriction does not hold or if another (Y,R,g) structure is deliberately chosen, then there are CAR situations that are R-CAR or not depending on the choice of R and $g(.,.)$. If this variable is just an addition to the original probability model, a randomisation device that only involves the distribution of (Y,Y^*) and that simply serves as a shortcut in computations involving the joined distribution, then there is no criterion to choose between competing possible versions and the choice can be based on the equivalence given above. However, more often than not, R is given some more or less intuitive interpretation. In that case, one must be aware that the statement of the CAR condition can not safely be formulated within the (Y,R,g) framework alone.

A particularly interesting case arises from censored data where the censoring is generally constructed from an extra variable R of censoring times such that

$$g(y,r) = \begin{cases} \{y\} & r > y \\ \{r,\ldots,\tau\} & r \leq y \end{cases} \tag{2.54}$$

so that the variable is censored if $R \leq Y$. Here, R is not a function of Y^*, but $g(.,.)$ at least satisfies (2.46) and (2.47). This last requirement also motivates the definition of censored sets by $r \leq y$ even though the

[47] Note that unions of strict refinements form a graded partially ordered set with respect to set inclusion in that all maximal linearly ordered subsets (flags) of $\mathcal{Y}^*$ have the same number of elements $|\mathcal{R}|$. Of course, unions of strict refinements are not the only possible forms of $\mathcal{Y}^*$. The requirement (2.53) says only that no element of $\mathcal{Y}^*$ is contained in more than one of the partitions $\mathcal{Y}^*_r$. A situation somewhat dual to that of strict refinements arises when no element of a partition $\mathcal{Y}^*_r$ is the union of elements from another partition. $\mathcal{Y}^* = \{\{1\},\{2,3\}\} \cup \{\{2\},\{1,3\}\}$ provides an example. Obviously, there are intermediate cases as well.

more traditional choice would take $r < y$. But with the latter convention, (2.54) would be violated when $y = r$. With this choice, if the model is assumed to be R-CAR, then the implied distribution of (Y, Y^*) must be CAR.

In the history of the analysis of censored data, the most prominent model certainly is that of a random censoring variable R that is independent of Y. The presumption is that there is a random variable R such that $R \perp\!\!\!\perp Y$ and such that $Y^* = g(Y, R)$. Within this setup, CAR follows since (Y, R, g) trivially satisfies (2.48), so it is R-CAR, and since $g(.,.)$ is of the right kind, it is also CAR. In consequence, the "simple device" sanctions the use of the face value likelihood that ignores how censored data come about.[48]

There are many drawbacks of this formulation that actually hampered the development of a clear understanding of censored data methods. A rather obvious problem occurs when the event of interest in a clinical study is the time of death of a patient. In that case, 'censoring after death' (the case $R > Y$) generally makes no sense when R refers to drop-out decisions or depends (possibly by protocol) on particular values of clinical variables. A possible remedy is to redefine R by setting $R \coloneqq \tau + 1$ if $|Y^*| = 1$ (Y is exactly observed) and $R \coloneqq \min Y^*$ otherwise. This convention does not change the status of R-CAR since R-CAR is trivially true when $|Y^*| = 1$. Moreover, the redefined R is a function of Y^* so that if this model is not R-CAR then it can not be CAR by the previous equivalence result. However, the redefinition destroys the global condition $R \perp\!\!\!\perp Y$, a condition that makes working with the (Y, R, g) model so attractive. In particular, if $Y^* = \{y\}$ then $R = \tau + 1$ so that independence can not hold.[49]

[48] Remember that the present framework is completely nonparametric so that problems connected with parameter dependence do not arise.

[49] See the remarks in van der Laan and Robins (2003: 23–24) as well as Rubin (2006) and the discussion of his article. Further discussion can be found in Egleston et al. (2007). A particular case is taken up by Frangakis et al. (2007) and Imai (2008). Those articles include references to arguments about counterfactual models and their connexion to the "simple device".
A superficially similar example with quite different consequences occurs when time to divorce (say Y) is of interest and death of a partner is treated as 'censoring'. This is just the reverse of the situation contemplated previously, where the value of an

As shown in the previous section, to any distribution of Y^* there is a joined distribution of (Y, Y^*) that is CAR. Since for censored data the set $\mathcal{Y}^*$ is hierarchically ordered, the solution of the CAR equations is straightforward. Moreover, the solution is unique on the finest partition with positive y^* probability. Of course, the latter two properties hold only when all elements of $\mathcal{Y}^* = \{\{0\}, \ldots, \{\tau\}, \{0, \ldots, \tau\}, \{1, \ldots, \tau\}, \ldots\}$ have positive probability. Note that Example 1 has the required structure except that other forms of $g(.,.)$ were considered.

Reformulating the example slightly by using $\mathcal{Y} = \{0, 1, 2\}$ instead of $\{1, 2, 3\}$ and giving all elements of $\mathcal{Y}^*$ the probability 1/5, the joined CAR distribution of (Y, Y^*) is:

	$\{0\}$	$\{1\}$	$\{2\}$	$\{1,2\}$	$\{0,1,2\}$	
0	1/5	0	0	0	2/40	1/4
1	0	1/5	0	1/10	3/40	3/8
2	0	0	1/5	1/10	3/40	3/8
	1/5	1/5	1/5	1/5	1/5	

In this case, it is easy to see that both distributions

	R	0	1	2	R'	0	1	2	
Y	0	2/40	1/15	2/15		2/40	4/40	4/40	1/4
	1	3/40	4/40	8/40		3/40	4/40	8/40	3/8
	2	3/40	4/40	8/40		3/40	4/40	8/40	3/8
		3/15	4/15	8/15		2/10	3/10	5/10	

imply the joined CAR distribution of (Y, Y^*) and that both are R-CAR. However, only in the left table are R and Y stochastically independent, while in the right table $\Pr(Y = 0, R' = 1) = 4/40 \neq \Pr(Y = 0)\Pr(R' = 1) = 1/4 \cdot 3/10$ and similarly for $\Pr(Y = 0, R' = 2)$. Here, R is a function of (Y, Y^*) on $\{Y \in \{1, 2\}\} \times \mathcal{Y}^* \cup \{Y = 0\} \times \{0, 1, 2\}$ so that only

assumed R was not well defined. It is now Y that is not a well defined quantity because it presupposes that a time to divorce Y is defined even if one of the partners died previously. But this is more an example of incompetent modelling than a problem of the random censoring model.

the probabilities $\Pr(R = 1 \mid Y = 0)$ and $\Pr(R = 2 \mid Y = 0)$ are undefined by the requirement that the joined distribution of (Y, Y^*) should be reproduced by the (Y, R, g) model. Hence, all feasible (Y, R, g) models are R-CAR in this situation, but some do not arise from independent censoring.

In fact, the situation presented above and its solution hold generally: When $g(.,.)$ is of the form (2.54), then to any $\mathcal{Y}^*$ representing censored data (i.e. when $\mathcal{Y}^*$ contains only sets of the form $\{y, \ldots, \tau\}$ and the singletons) and any distribution of Y^* there are random variables R such that (Y, R, g) is R-CAR and implies the CAR distribution of (Y, Y^*). To see this, one has only to note that CAR and R-CAR are restrictions only for the censored sets in $\mathcal{Y}^*$. But for these (and $g(.,.)$ of the given form), R is a function of Y^* so that the argument for strictly invertible $g(.,.)$ applies here as well. The general condition of invertibility (2.53) is too strong. It is enough to require that r is a function of y^* only for those y^* with $|y^*| > 1$.

Furthermore, the conditional independence of these R-CAR models implied by (2.51) can be extended to unconditional independence (just counting equations) and this extension defines an independent censoring model (i.e. $R \perp\!\!\!\perp Y$). Thus to any censored data model there is one that can be represented by an independent censoring variable R.

This simple fact, while known for special cases (e.g. Berman 1963), was in some generality only observed in 1977 by Williams and Lagakos. In particular, they showed that a CAR distribution of (Y, Y^*) (in the absolutely continuous case) can always be represented by an independent censoring model (Williams and Lagakos 1977: Theorem 3.1). Their constant sum condition has prompted the development of quite a few reformulations of conditions where likelihood inference based on the face value observations is valid (Kalbfleich/MacKay 1979, Ebrahimi et al. 2003). The existence of several R-CAR distributions in the (Y, R, g) model was also noted for interval censoring (when only a random interval is known during which the event of interest happened) and for current status data (when it is only known whether an event happened before a random inspection time or not). See Oller et al. (2004, 2007) and Lawless (2004) for the case of interval censoring and Betensky (2000) for current status models. For competing risks, Langberg et al. (1978) noted the usefulness of constructing independent representations of the type (Y, R, g) for the handling of many questions in the description of stochastic systems. But only much later, Crowder (1996, 1997, 2000) clarified the situation also for discrete and mixed time models.

The relation of the (Y, R, g) formulation to CAR becomes even more complex when multidimensional censored times are considered. In this case, one may consider a $g(.,.)$ as in (2.54) applied coordinate-wise: If (Y_1, Y_2) and (R_1, R_2) are times and censoring times, then

$$g(y_1, y_2, r_1, r_2) = \begin{cases} \{(y_1, y_2)\} & r_1 > y_1, r_2 > y_2 \\ \{r_1, \ldots, \tau_1\} \times \{y_2\} & r_1 \leq y_1, r_2 > y_2 \\ \{y_1\} \times \{r_2, \ldots, \tau_2\} & r_1 > y_1, r_2 \leq y_2 \\ \{r_1, \ldots, \tau_1\} \times \{r_2, \ldots, \tau_2\} & r_1 \leq y_1, r_2 \leq y_2 \end{cases} \tag{2.55}$$

But now it may happen that a CAR distribution can not be represented by independent pairs $(Y_1, Y_2) \perp\!\!\!\perp (R_1, R_2)$. For an example, suppose that (Y_1, Y_2) are times to consecutive events so that the event times are $(Y_1, Y_1 + Y_2)$. Let R be an independent variable that censors $(Y_1, Y_1 + Y_2)$ in calendar time. Then either $R > Y_1 + Y_2$ so that $Y^* = (Y_1, Y_1 + Y_2)$, or $Y_1 < R \leq Y_1 + Y_2$ so that $Y^* = \{Y_1\} \times \{R, \ldots, \tau\}$, or, lastly, $R \leq Y_1$ so that $Y^* = \{R, \ldots, \tau\} \times \{R, \ldots, \tau\}$. When $g(.,.)$ is chosen as in (2.55), then one may say that (R, R) censors $(Y_1, Y_1 + Y_2)$. This model is trivially R-CAR and since r is a function of y^* for the censored elements of $\mathcal{Y}^*$ it is also CAR.

If one is interested in the joined distribution of (Y_1, Y_2) (each episode starts from time 0), then the correspondingly transformed $\mathcal{Y}^*$ consists of either complete observations (y_1, y_2), or the half lines $\{y_1\} \times \{y_2, \ldots, \tau_2\}$, or the half spaces $\{y_1, \ldots, \tau_1\} \times \mathcal{Y}_2$. The form of incomplete elements of $\mathcal{Y}^*$ is illustrated in Figure 2.13.

The model clearly is still CAR. After all, it is just a one-to-one transform of the original model. If one constructs an (R_1, R_2) for the $g(.,.)$ in (2.55), then one gets $R_1 := R, R_2 := \max(0, R - Y_1)$. But if $(Y_1, Y_2) \perp\!\!\!\perp R$, then

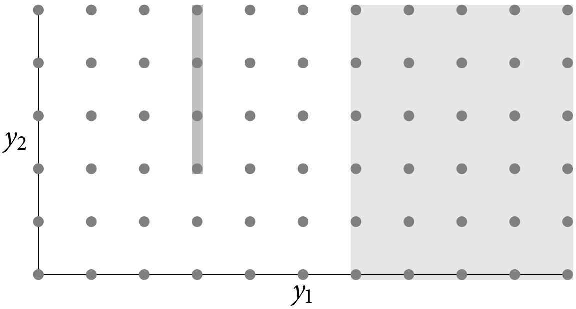


Figure 2.13.: The possible forms of coarsened regions in a two-dimensional censored data model.

this can not be the case for (Y_1, Y_2) and (R_1, R_2): R_2 is a function of Y_1 unless $\Pr(\{R < Y_1\} \cup \{R > Y_1 + Y_2\}) = 1$.

It follows that there are CAR models that can not be represented by independent censoring variables (R_1, R_2). Modelling bivariate censored data by an independent censoring model (Y, R, g) restricts the class of CAR models. Gill et al. (1997: 270–272) note the statistical consequences: When only CAR is assumed, then one must use the non-parametric maximum likelihood estimator (see also Chapter 8). On the other hand, when only the smaller set of models that assume independent censoring according to (2.55) is contemplated, then there are competing, non-equivalent estimators that may be even better (or at least more stable) than the non-parametric maximum likelihood estimator in small samples.

Moreover, R_1 is not a function of y^* when y^* is the half line $\{y_1\} \times \{r_2, \ldots, \tau_2\}$. One may therefore change the conditional distribution of R_1 given Y on $\{R_1 > y_1\}$ so that it depends on the values of $y_2 \in \{r_2, \ldots, \tau_2\}$ without changing the joined distribution of (Y, Y^*). Thus there can be (Y, R, g) models that are not R-CAR according to (2.48) but that imply a joined distribution of (Y, Y^*) that is CAR. R-CAR is not a necessary condition for CAR. Working within a (Y, R, g) model for multidimensional censored data makes it difficult to decide whether the CAR condition can be stipulated, even for rather restricted structures of

$\mathcal{Y}^*$.

2.3.2. Completely Coarsening at Random

A further consideration in connection with the (Y, R, g) model of considerable recent interest is a strengthening of the R-CAR condition. It is a condition that has been termed *coarsening completely at random* (CCAR). It is a rather obvious extension in the case of missing data where it simply states that incompleteness does not depend on any other information in the model. The standard example is that of a missing dependent variable in a regression model where CCAR says that the probability of a missing dependent variable does not depend in any way on the values of the covariates. The extension to general incomplete data models was suggested by Heitjan (1994).

The CCAR condition can be easily defined in terms of (Y, R, g) models by strengthening condition (2.48) to

$$\Pr(R = r \mid Y = y) \text{ is constant} \tag{2.56}$$

or equivalently—strengthening condition (2.51)—as

$$\{Y = y\} \perp\!\!\!\perp R \tag{2.57}$$

for y of positive probability. When CCAR is defined using the (Y, R, g) model I call any one of the conditions the R-CCAR condition.

The condition (2.57) asking for stochastic independence of Y and R is arguably the most famous notion of CAR cited in the applied literature. The fame of this strong version of R-CAR stems from the fact that far reaching conclusions can be drawn from it with ease.

In view of the difficulties of R-CAR conditions to provide an unanimous definition of CAR it would be very convenient to have a formulation that could be based on the distribution of (Y, Y^*) without recourse to some R. It turns out that this is an extremely difficult task. A little contemplation reveals that for a strengthening of the condition (2.28) one would need to require

$$\Pr(Y^* = y^* \mid Y = y) \text{ is constant} \tag{2.58}$$

which is self-contradictory unless the variable Y^* would only take the value $\mathcal{Y}$ so that the incomplete variable carries no information at all. The reason is that for some y and an y^* not containing y the probability must be 0. It follows that the probability is 0 for all y and thus the marginal probability of any $y^* \neq \mathcal{Y}$ must be 0. The same problem shows up in all the alternative formulations of the CAR conditions.

There were a few attempts to remedy the situation. An obvious one was suggested by Jaeger (2005b) who in parallel with the CAR situation defined a model distribution to be CCAR if there was at least one R-CCAR version of (Y, R, g) compatible with (Y, Y^*).

I will argue here that CCAR is properly viewed as a condition on the structure of $\mathcal{Y}^*$ that extends CAR and not as a stronger version of the CAR condition itself. If CAR is equivalent to a factorisation of the fully non-parametric incomplete data likelihood as in (2.32) then there can be no further information that would ameliorate the estimation problem. CAR without any further conditions permits to use only the variable of interest part of the likelihood while disregarding the coarsening model, at least for likelihood inference. In fact, most examples for R-CCAR in contrast to CAR introduce further variables into a problem and equate R-CCAR then with the ability to drop the relation of these extra variables to the variable R and thus the relevant likelihood factors from consideration. But this is more like a re-expression of the standard CAR condition extended to more complicated models. Examples of R-CCAR that are not just barely disguised versions of CAR hint in another direction: Loosely speaking, R indicates the degree of incompleteness. Thus the R-CCAR condition says that it is possible to determine the degree of incompleteness without knowledge of the actual y drawn in a procedure. Therefore, CCAR should single out a certain class of procedures to generate incomplete data. These procedures should not depend in any way on the procedures used to create the values of Y. If that was the case, then the outcome of combining the independent procedures to create the coarsening and to create the variable of interest would be a special case of CAR, or so one would hope.

Such independent coarsening procedures certainly exist: If $\mathcal{Y}^*$ is a union of partitions $\mathcal{P}_1, \ldots, \mathcal{P}_k$ one can do the following: select one of the partitions with probability p^r. Next generate an y according to the marginal distribution of Y and report as y^* the element of the chosen partition into which y falls. As Jaeger (2005b: Theorem 2.14) shows, this is also the only way R-CCAR may come about: It follows from the restrictions (2.46) and (2.47) on $g(.,.)$ already discussed in the previous subsection. As shown there, any R-CAR model (Y, R, g) where $g(.,.)$ satisfies (2.46) and (2.47) implies that $\mathcal{Y}^*$ is a union of partitions indexed by $r \in \mathcal{R}$. If the model is R-CCAR as well, then R and therefore the partition chosen is independent of Y and so the above algorithm applies. In the other direction, one needs only to set $\mathcal{R} = \{1, \ldots, k\}$ and $\Pr(R = r \mid Y = y) = p^r$ for all y of positive probability. This is obviously R-CCAR.

2.3.3. Coarsening Variables as Additional Data

Even though the (Y, R, g) model and the R-CAR conditions can only be made equivalent with CAR under very restrictive conditions on $\mathcal{Y}^*$ and the form of $g(.,.)$ and $\mathcal{R}$, there is a sense in which it might be more general than the model solely based on (Y, Y^*). Consider the case of censored data represented by a censoring time R. Occasionally it is possible to identify such a variable with further information in the situation. This might be the case when all censoring events are recorded even when they happen after the event of interest. In that case the value of R would always be known and should be introduced into the model. This can not be represented in terms of the original variables (Y, Y^*) alone. One must at least expand the set of contemplated observations to some $\mathcal{Z}$, say. Then CAR and similar concepts must relate to the common distribution of (Y, Z) where Z is a random variable with values in $\mathcal{Z}$. To connect this general setup with models of incomplete data, there must be a function

$$\alpha \colon \mathcal{Z} \longrightarrow \mathcal{P}(\mathcal{Y}) \tag{2.59}$$

that presents the coarsened values of Y such that consistency holds, i.e.

$$\Pr(Y \in \alpha(Z)) = 1$$

The generalisation fits well with the (Y, R, g) formulation of the previous section. Now one would set

$$g \colon \mathcal{Y} \times \mathcal{R} \longrightarrow \mathcal{Z}$$

and put

$$\alpha(z) \coloneqq \{y \in \mathcal{Y} \,|\, \exists r \in \mathcal{R} : g(y, r) = z\}$$

An obvious advantage of using a general space $\mathcal{Z}$ together with a function $\alpha(.)$ instead of $\mathcal{Y}^*$ is the ability to use mathematically convenient spaces in place of sets of subsets. In the censored data case I have already used informally the standard notation (Y', δ). The corresponding $\alpha(.)$ is the one-to-one transformation with $\alpha(y, 1) = \{y\}$ and $\alpha(y, 0) = \{y, \ldots, \tau\}$. Furthermore, the move allows to work with situations where the set of subsets is too large to allow a reasonable probabilistic treatment. An example is the arbitrary grouping of a continuous variable to subsets of the real line. If the subsets can be taken to be intervals, one can replace the sets by their upper and lower bounds so that one can conveniently work in $\mathbb{R}^2$.[50]

Mathematical convenience aside, the impact of additional 'information' on the concept of CAR is of primary interest here.[51] A direct generalisation of the CAR conditions (2.21), (2.28), and (2.30) gives either

$$\Pr(Y = y \,|\, Z = z) = \Pr(Y = y \,|\, Y \in \alpha(z)) \tag{2.60}$$

[50] The translation into manageable random variables Z was used by Gill et al. (1997: Sections 6–9) to formulate CAR when $\mathcal{Y}$ is infinite, and in particular when $\mathcal{Y} = \mathbb{R}$. The most appropriate way to do so is still debated. See Jacobsen and Keiding (1995), Nielsen (2000), and Cator (2004) for some critical remarks.

[51] The term 'information' is used in quotes here, since random variables and conditional distributions have properties that sometimes contradict the properties the term carries in informal use. See Dubra and Echenique (2004) for an example similar to Billingsley's (1979: Example 33.11).

or

$$\Pr(Z = z \mid Y = y) \text{ is constant on } y \in \alpha(z) \tag{2.61}$$

or

$$\{Y = y\} \perp\!\!\!\perp \{Z = z\} \mid \{Y \in \alpha(z)\} \tag{2.62}$$

for y of positive probability. The equivalence of these conditions follows as in the previous, simple case. I will call this generalised CAR condition *augmented* CAR or *A-CAR*.

From the last condition, since Y^* is a (measurable) function $\alpha(.)$ of Z, the simple CAR condition (2.30) ($\{Y = y\} \perp\!\!\!\perp \{Y^* = y^*\} | \{Y \in y^*\}$) follows from the condition (2.62). The requirements (2.60)–(2.62) can therefore be strictly stronger than the CAR condition discussed thus far. Consequently, one can not expect that the slogan 'CAR is everything' will hold in this situation. There may well be distributions of Z (and functions $\alpha(.)$) such that no joined distribution of (Y, Z) satisfying (2.60)–(2.62) exists. It is clear that adjoining further 'information' to Y^*— whether in the form of an explicit random variable R in a generalised model (Y, R, g) or simply as a joined distribution of (Y, Z)—will generally destroy the conditional independence relation that can always be established for simple CAR. After all, knowing $\{Z = z\}$ might entail much more detailed 'information' than knowing $\{Y^* = \alpha(z)\}$.

The Monty Hall problem illustrates the expressiveness of A-CAR: Until now I took as starting point the moment when the contestant has announced his choice of door 2. In that case one might take $\mathcal{Y} = \{1, 2, 3\}$ and $Y^* = \{\{1, 2\}, \{2, 3\}\}$ as before. Now suppose that we step back to the situation just before the contestant chooses a door. Then there is a new element that needs to be represented in the model: The initial choice of the door number by the contestant. Denote this choice by C. Furthermore, denote by R (the number of) the remaining door, the one *not* opened by Monty and *not* chosen by the contestant (this is known after the contestants choice and Monty's reaction). Then, according to the rules of the game, the price must be either behind door C or behind door R. One may define $Z \coloneqq (R, C)$ as the relevant information.

After all, both the choice of the contestant and the choice of the show master might be regarded as part of the probability model describing the Monty Hall problem. With both parts of 'information' available—the choice of the contestant and the choice of the host—one should be able to represent the outcome as $\{(C = c, R = r)\}$. The function $\alpha(.)$ maps the ordered pair to the set $\{R, C\}$ with the same interpretation as before: The doors that may hide the price after Monty opened the door with the goat. Suppose that C, the choice of the contestant, is independent from Y. Now

$$\Pr(R = 1 \mid Y = 1, C = 2) = 1$$

since Monty has no choice when the price is not behind the door chosen by the contestant. If he has a choice, suppose it is not deterministic so that

$$\Pr(R = 1 \mid Y = 2, C = 2) =: p_1 \notin \{0, 1\}$$

Then

$$\begin{aligned}\Pr(Y = 1 \mid Z = (1, 2)) &= \frac{\Pr(C = 2)\Pr(Y = 1)}{\sum_y \Pr(Z = (1, 2) \mid Y = y)\Pr(Y = y)} \\ &= \frac{\Pr(Y = 1)}{\Pr(Y = 1) + p_1 \Pr(Y = 2)} \\ &\neq \Pr(Y = 1 \mid Y \in \{1, 2\}) \\ &= \Pr(Y = 1 \mid Y \in \alpha((1, 2)))\end{aligned}$$

so that (2.60) is wrong unless $\Pr(Y = 2) = 0$ or $\Pr(Y = 1) = 0$. By a similar computation, $\Pr(Y = 3 \mid Z = (3, 2)) \neq \Pr(Y = 3 \mid Y \in \{2, 3\})$ unless $\Pr(Y = 3) = 0$ or $\Pr(Y = 2) = 0$. Thus on $\{C = 2\}$ A-CAR can only hold when $\Pr(Y = 2) \in \{0, 1\}$. This is of course the same result as the one obtained earlier. After all, the derivation was basically conditional on $\{C = 2\}$. But the previous discussion could only proceed unambiguously by conditioning on the choice of the contestant. Otherwise, using the possible locations of the price as the coarsened information could be misleading: The set $\{1, 2\}$ might come about because the contestant has chosen door 2 and Monty had opened door 3, or because the contestant had chosen door 1 and Monty opened door 3, a difference that can not be expressed in the subset formulation. With A-CAR, we can now speculate explicitly about both C, the choice of the contestant, and R, Monty's choice.

First, if one considers different choices of the contestant, the CAR (or A-CAR) condition requires $\Pr(Y = 1) \in \{0, 1\}$ if $C = 1$, $\Pr(Y = 2) \in \{0, 1\}$ if $C = 2$, and $\Pr(Y = 3) \in \{0, 1\}$ if $C = 3$. If at least two of the three choices have positive probability, then the second CAR solution ($\Pr(Y = 2) = 0$ and $\Pr(Y = 1) = \Pr(Y^* = \{1, 2\}), \Pr(Y = 3) = \Pr(Y^* = \{2, 3\})$) is ruled out. Note that I still use the weak version of CAR where the condition is trivially true for degenerate distributions of Y.

Secondly, one may consider Monty's behaviour. The above argument worked only if he decided non-deterministically. Now suppose he always chooses the larger number if he has a choice. So he opens door 3 (and never door 1) when the contestant has chosen door 2 and the price is behind that door. Then

$$\Pr(R = 1 \mid Y = 2, C = 2) = 1$$

and therefore

$$\begin{aligned}\Pr(Y = 1 \mid Z = (1, 2)) &= \frac{\Pr(C = 2)\Pr(Y = 1)}{\sum_y \Pr(Z = (1, 2) \mid Y = y)\Pr(Y = y)} \\ &= \frac{\Pr(Y = 1)}{\Pr(Y = 1) + \Pr(Y = 2)} \\ &= \Pr(Y = 1 \mid Y \in \alpha((1, 2)))\end{aligned}$$

in accordance with A-CAR. But now

$$\Pr(R = 3 \mid Y = 2, C = 2) = 0$$

so that if he does open door 1 and thus $R = 3$, the price must be behind door 3,

$$\Pr(Y = 3 \mid Z = (3, 2)) = 1 \neq \frac{\Pr(Y = 3)}{\Pr(Y = 2) + \Pr(Y = 3)}$$

unless $\Pr(Y = 2) = 0$ (or $\Pr(Y = 3) = 0$, in which case the condition is irrelevant), a weaker condition then the previous result. Such possibilities are generally overlooked when the Monty Hall problem is discussed using some kind of simulation.

While A-CAR can be much more expressive and easier to work with, it is not obvious what the role of *R* should be. If it represents additional 'information', then presumably it should be incorporated into the model by adding it to Y. Here it should make no difference whether *R* is completely specified (*R* is 'observable' in Jaeger's terminology) or only partial given as in the (Y, R, g) formulation. In this perspective, A-CAR is not a generalisation of CAR. The introduction of *R* or other auxiliary variables is simply an indication that the original model was chosen too small. In the properly enlarged model, A-CAR reduces to CAR.

On the other hand, such an argument disregards the requirement that *Y* ought to represent particular facts within the model. When *Y* and $\mathcal{Y}$ are used to express the social facts of interest, extensions of the definition of *Y* can not be justified by particular considerations of aspects of the "simple device". But now the additional random variable *R* certainly introduces further structure beyond the one present in the simple probability model, at least when *R* is not just a function of a possibly extended Y^*. And as was demonstrated in the last section, the corresponding R-CAR and A-CAR conditions need not be equivalent with CAR. In particular, when *R* is used to represent genuine further 'information' that nevertheless should not be incorporated into a larger model, then A-CAR will often fail. The slogan 'A-CAR is everything' is certainly not correct. A-CAR can be rejected empirically. It can not be invoked in the same way as CAR can, as a simple extension of the "simple device". It is not a modelling decision.

Note, however, that the decision to distinguish between *Y* and *R* is a modelling decision. It will almost always rest on an idea of a 'mechanism' that brings about incompleteness. *R* is used to represent aspects of this 'mechanism' and, since the 'mechanism' is thought to operate after the fact represented by *Y* came about, it can not be part of *Y* itself. While I do not think that the idea of a 'mechanism' acting independently from *Y* can be given a reasonable interpretation in social science applications, the distinction it hints to might still be relevant when versions of A-CAR are contemplated.

2.4. Conclusions: CAR Modelling

Even though the connection with empirical investigations still needs clarification and the discussion up to now was confined to the model world, the results for the "simple device" are much stronger than could be expected from a discussion within the framework of classical sampling theory. First of all, I have shown that the introduction of a probability model does not restrict the possible underlying values beyond the consistency requirement so that the "simple device" does not prejudice certain solutions against others. Secondly, it is possible to show the mathematical equivalence of several formalisations of randomly coarsened data. This can't be done in the sampling framework since in that case decompositions and independence relations can hold at most approximately. And it is possible to deduce that the distribution of randomly coarsened variables will imply a unique distribution of the variable of interest. The latter result is more or less obvious in the case of either exact or completely missing data where it may be formulated within the classical framework of sampling as well. But in all other cases, and thus in situations of much larger practical relevance, this could only be done within the probability model.

Since to any distribution of the coarsened variables there always exists a (weak) CAR model, it is impossible to reject the suggestion by referring to observations of the coarsened variables. CAR is always a viable choice in the model framework that can not be criticised on empirical grounds. CAR can therefore serve as a reference point for further speculations about the modelling within the probabilistic framework. Such speculations are not arbitrary. They are constrained by the very construction of probability models. But they clearly are not assumptions about reality. They can definitely only be formulated within the probability model. They thus presuppose the "simple device".

It is however possible to judge the merits of such a model by appeal to other principles. The existence of a 'mechanism' may serve to exclude some cases where CAR would seemingly imply more than the validity of taking the coarsened variables at face value. When a 'mechanism' in the strong sense of a well defined transition kernel from $\mathcal{Y}$ to $\mathcal{Y}^*$ does not seem to be appropriate, one must be cautious when CAR implies several (degenerate) solutions. Similarly, extensions of CAR like R-CAR and A-CAR introduce further elements that might be used to judge model adequacy. But it turns out that these extensions can not be employed as a general reference point in the same way as simple CAR does.

While it is reassuring to learn that the "simple device" works without prejudicing certain solutions, it must be remembered that the critics refer to answer patterns of real people in real surveys. The theoretical analysis provided here can do no more than to show that the "simple device" is not inadequate by construction. An answer to the concerns of the critics, however, will also need a critically appraisal of empirical work.

3

A Case Study: Parent's Length of Life

This Chapter studies a particular case of incomplete data. It does not aim to provide an appraisal of empirical work using the "simple device". Rather, it provides a view of the working of the "simple device" in a complicated setting, a setting not covered by the now classical applications to missing, grouped, truncated, or censored data. Moreover, the relation of the "simple device" to classical survey statistics is examined.

The practical problem posed here is how to gain insight into mortality conditions during the early 20th century from current survey data. While official statistical agencies routinely provide life tables since the late 19th centuries, these are naturally restricted in scope and detail. Survey data may provide more background information and apply also to times and geographical areas not covered by the official statistics. More precisely, additional information about mortality in the early 20th century can be gained from current surveys when respondents were asked to provide information about their parents, in particular about their parent's birth years, whether they were still alive at the interview date, and, if not, about their respective years of death. Both the German Life History Study (GLHS) and the Socio-economic Panel (SOEP) provide such information and one might try to use this type of information to enlarge the knowledge about mortality conditions in earlier periods.[1]

[1] For previous analyses of the SOEP data about the life lengths of parents see Schepers

On the other hand, survey data on mortality of the parent generation naturally suffer from incompleteness for various reasons. Since the information about parents is supplied by their children, life length of childless persons is unavailable. It is not just missing: From survey data alone even the number of such persons can not be estimated. Moreover, the information is only available when the children themselves survived up to the survey date. And the chance of this to happen will also depend on the number of children of the parents. Thus, even though the focus is on the mortality of the parent generation, the mortality of the current generation must be taken into account. A further complication arises from the possibility that the same person is included several times in the survey because several of her or his children were questioned. This will in general be difficult or impossible to check using survey methodology. Relatedly, people with many children have a higher probability to be included in the survey. Lastly, it should not be forgotten that migration and expulsion as well as changes of territorial borders make it difficult to maintain the idea of a uniquely defined target population. In this Chapter, however, I will cling to the idea in order not to overburden the discussion.

3.1. The Structure of Descent

The first problem to be treated is that of possible dependence by descent and of unequal inclusion probabilities. In this Section and the next, I will abstract from all other aspects of the data and consider only what this type of incompleteness implies for the analysis strategies.

An extremely simplified model of the situation is depicted in Figure 3.1. Members of a generation are represented on horizontal lines by circles

and Wagner (1989), and Klein (1993). In some cases, survey data about about life lengths of relatives or household members are the only source of information on mortality available. See Gakidou and King (2006) for further references to applications and a discussion of of standard methods.

(for women) and squares (for men).[2] The generations follow each other through time t = -2, -1, 0 down the Figure. Arrows point from parents to children. Thus all generations are exactly aligned, there is no overlap of generations. Moreover, members of a generation are distinguished by sex and the number of their offsprings only. Life length will be thought of as an additional attribute of the nodes of the graph, definable without regard to either the placement into a certain generation or the number of offsprings.

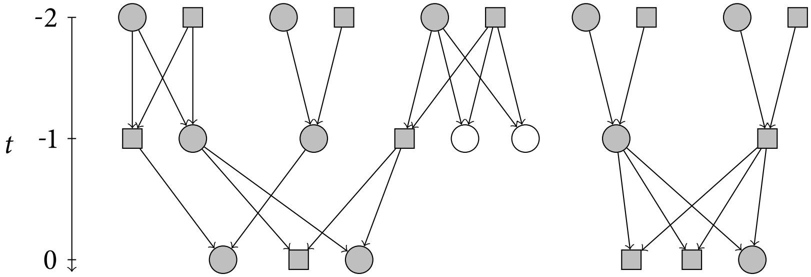


Figure 3.1.: A complete ancestry graph. Time runs from the oldest generation at top down to the youngest. Men and Women are distinguished as circles and squares. Looking backwards in time, information on the two empty circles is unavailable.

Looking forward in time, starting at t = -2 and following the arrows down the Figure allows to trace out all ancestral dependencies between members of different generations. I will call the graph representing the complete ancestral dependencies the *ancestral graph*.[3] However, looking backward in time, starting at t = 0 and following the arrows in the opposite direction will recover only a part of the ancestral graph. In the Figure, the two empty circles can not be reached from any member of the population at time t = 0. The resulting graph will be called *pedigree*

[2] It should be obvious that the terms 'generation' as well as 'parent', 'child', 'men', etc. pertain to the model entities presented here. I have therefore abstained from making additional typographical distinctions.

[3] This use of the word "ancestral graph" should not be confused with the ancestral graphs used in the theory of graphical models. See Richardson and Spirtes (2002, 2003) for the latter concept.

graph.[4]

Suppose the members of a generation can be listed in the form

$$\mathcal{U}_t \coloneqq \{u_{t,1}, \ldots, u_{t,N_t}\}$$

so that the size of a generation is $|\mathcal{U}_t| = N_t$. Also, let the generations be partitioned into their male and female members so that

$$\mathcal{U}_t = \mathcal{U}_t^f \cup \mathcal{U}_t^m \text{ and } \mathcal{U}_t^f \cap \mathcal{U}_t^m = \emptyset$$

Further write $\mathcal{U} \coloneqq \cup_t \mathcal{U}_t$ for the set of members of all generations considered. Then the ancestral graph is technically speaking a directed, acyclical, bipartite graph whose nodes are the elements of $\mathcal{U}^f \cup \mathcal{U}^m$. Translating the conditions of an ancestral graph into the language of graph theory, all nodes of positive indegree must have indegree exactly 2 (all people have two parents, up to the oldest generation considered). Furthermore, if there is an edge $(u_{t,i}, u_{t+1})$ from $u_{t,i}$ to one u_{t+1} and a different edge $(u_{t,j}, u_{u+1})$ to u_{t+1}, then either $u_{t,i} \in \mathcal{U}_t^f$ and $u_{t,j} \in \mathcal{U}_t^m$, or $u_{t,i} \in \mathcal{U}_t^m$ and $u_{t,j} \in \mathcal{U}_t^f$ (all people have a mother and a father). Note that the subgraphs that arise from deleting either all nodes $\mathcal{U}^f$ or all nodes $\mathcal{U}^m$ and the respective edges gives a *forest*, a union of trees. Thus, retricting attention to either fathers or mothers simplifies analyses considerably.

In order to deal with the ancestral information from a statistical point of view, define a pair of statistical variables that record the mother and father of each member of a generation as

$$\begin{aligned} m &: \mathcal{U}_{t+1} \longrightarrow \mathcal{U}_t^f \\ f &: \mathcal{U}_{t+1} \longrightarrow \mathcal{U}_t^m \end{aligned}$$

[4] The reconstruction of pedigree or ancestral graphs from partial information is an important topic in biology. There is a large specialised literature dealing with superficially similar problems. See Tavaré (2004) for a vivid account of probability models for pedigree graphs and Steel and Hein (2006) for a particular application. However, the focus of that literature is on many generations and rarely touches statistical problems. There is also a formal connexion to the theory of random refinements of intervals (see e.g. Bertoin 2006), which, however, also concentrates on long sequences.

giving the mothers (elements of $\mathcal{U}_t^f$) and fathers (elements of $\mathcal{U}_t^m$) of all members of the generation $\mathcal{U}_{t+1}$. This is a well defined function since everyone has both a father and a mother. Now the set of parents can be written as

$$\mathcal{U}_{\text{-}1}^p \coloneqq \{m(u_0) \mid u_0 \in \mathcal{U}_0\} \cup \{f(u_0) \mid u_0 \in \mathcal{U}_0\}$$

and one of the problems to be discussed can now be formulated as

$$\mathcal{U}_{\text{-}1}^p \subsetneq \mathcal{U}_{\text{-}1}$$

i.e., the *parent generation* $\mathcal{U}_{\text{-}1}^p$ is in general a strict subset of the *previous generation* $\mathcal{U}_{\text{-}1}$.

To make reference to children easier, it is convenient to introduce the set valued function

$$\begin{aligned} c\colon \mathcal{U}_t &\longrightarrow \mathcal{P}(\mathcal{U}_{t+1}) \\ c(u_t) &\coloneqq \begin{cases} m^{\text{-}1}(\{u_t\}) & \text{if } u_t \in \mathcal{U}_t^f \\ f^{\text{-}1}(\{u_t\}) & \text{if } u_t \in \mathcal{U}_t^m \end{cases} \end{aligned}$$

Thus, $c(u_t)$ is just the set of children of u_t and the set of parents can also be written as $\mathcal{U}_{\text{-}1}^p = \{u_{\text{-}1} \in \mathcal{U}_{\text{-}1} \mid c(u_{\text{-}1}) \neq \emptyset\}$.

To complete the description of this simplified model, a variable representing life length is attributed to each member of the population:

$$T\colon \mathcal{U} \longrightarrow \mathcal{T} \coloneqq \{0, 1, \ldots, \tau\}$$

where in accordance with the discussion in Chapter 2 life length is supposed to take a finite number of values. Note that I here do not impose any restrictions on $T(.)$, it may take any value. In particular, $T(u) = 0$ and $|c(u)| > 0$ is not ruled out. Also, $T(u)$ is assumed to be defined for all u, even for the currently living generation.

Finally, in the following I will drop the reference to the generation subscript whenever there is no danger of confusion. In particular, $\mathcal{U}_{\text{-}1}^p$ will be denoted by $\mathcal{U}^p$ and $\mathcal{U}^c \coloneqq \mathcal{U}_0$ will denote the set of members of the current generation, the children of $\mathcal{U}^p$. Similarly, subscripts to the elements of these sets will be dropped when possible.

3.2. Sampling from the Children Generation

In this Section, I will ignore the problem that in general $\mathcal{U}^p \subsetneq \mathcal{U}$. Instead, I will pretend that interest actually centres on the parent generation. This redefines the original question by excluding childless persons from the target population. Then only two related difficulties remain: First, the more children someone of the parent generation has, the higher the probabilty of obtaining information about her or his life length. Secondly, he or she may be included as many times in the survey as she has children, but this will generally be unknown to the survey statistician.

To discuss the impact of these features on the sampling approach, let a *sample* be any subset of the population $\mathcal{U}^c$. A collection of samples $\mathcal{S} := \{s_1, s_2, \ldots\}$ together with a probability distribution defined on the elements of $\mathcal{S}$ is called a *design*. Furthermore, let S denote a random variable on some probability space, say $(\Omega^S, \mathcal{A}, \Pr)$, with values in $\mathcal{S}$ whose distribution equals the design distribution. Thus

$$\begin{aligned} S\colon\ \Omega^S &\longrightarrow \mathcal{S} = \{s_1, s_2, \ldots\} \\ \Pr(S = s) &:= \Pr(\{\omega \in \Omega^S \mid S(\omega) = s\}) \end{aligned}$$

where measurability is taken for granted. In order to distinguish this probability from other distributions—specifically those used by the "simple device"—I will write $p(s)$ for $\Pr(S = s)$. Suppose for simplicity that the design is that of a simple random sample without replacement, i.e. all samples $s \in \mathcal{S}$ are of equal size $|s| =: n$, say, and all samples have the same probability. Then $\Pr(S = s) = p(s) = 1/\binom{N_0}{n}$ and the inclusion probabilities of the first and second order are

$$\begin{aligned} \pi(u) &:= \Pr(S \ni u) = \sum_{s \in \mathcal{S}} \mathbb{1}[s](u) p(s) = \frac{n}{N_0} \\ \pi(u, u') &:= \Pr(S \ni u \cap S \ni u') = \frac{n(n-1)}{N_0(N_0 - 1)} \quad \text{for } u \neq u' \end{aligned}$$

These inclusion probabilities are the backbone on which most computations in the classical sampling theory rest.

When a sample is chosen according to this probability distribution (using some appropriate random device), one proceeds to ascertain the the life lengths of the parents of all sampled persons, i.e. $T(m(u))$ and $T(f(u))$ for all $u \in S$.[5]

3.3. Induced Sampling from the Parent Population

A natural next step is to consider $\{m(u) \mid u \in s\} \cup \{f(u) \mid u \in s\}$, the fathers and mothers of the people included in the sample s, as a sample from $\mathcal{U}^p$. This is superficially similar to sampling with replacement in that parents of several children may be included several times. The multiplicities of inclusion do not, however, result from a sequential drawing from the complete population as in most sampling designs with replacements.

I will denote the *set* of sampled parents by s^*. Thus,

$$s^* \coloneqq \{m(u) \mid u \in s\} \cup \{f(u) \mid u \in s\}$$

which I will call the *set-sample* from $\mathcal{U}^p$. The multiplicities of elements u of s^* will be denoted by $N(\{u\})$. The set s^* together with the map $u \mapsto N(\{u\})$ for $u \in s^*$ will be called a multiset, denoted by s^{**}.[6]

[5] Textbooks more often than not liken the values of sample statistics to random variables, possibly adding even stronger qualifications like independence or identical distributions. The reasoning seems to be that even if $T(u)$ is a fixed quantity for each u, whether it is included in a sample will only depend on the sampling design. In consequence, one might conceive of a random sample as a probability sample from $\mathcal{T}$, or so the reasoning goes.
But on closer inspection, the reference should be to the function $T^* : \Omega^S \rightarrow \mathcal{T}^n$ which is the composition of T and S. Since $\mathcal{T}$ is finite and $S(.)$ is assumed to be measurable, the function T^* is in fact a well defined random variable. But since the values of this function will be unknown for all $\omega \in \Omega^S$ except for the one used in the actual sample, nothing more can be said about this function. The use of the definite article when referring to such a function will be easily misleading. It seems much better to think of $T^*(.)$ as a whole class of random variables.

[6] See Stanley (1997: Chap. 1.2) for some conventions about multisets. The multiset s^{**} is called an *ordered sample* by Särndal et al. (1992: 49) because they consider sequential draws from a population with replacement. Since this is not what actually

An example may make the situation clear. Suppose the following pedigree graph is given: Suppose further that a sample of size 3 is drawn without replacement from the 5 children. Then a possible sample comprises only the children of the first couple, $s = \{u_{0,1}, u_{0,2}, u_{0,3}\}$ and this induces a sample $s^{**} = \{u^3_{-1,1}, u^3_{-1,2}\}$ of the parent generation, where the multiplicities are noted as superscripts. This is obviously the sample of minimal *effective sample size* $|s^*| = 2$. If the sample is $s = \{u_{0,1}, u_{0,4}, u_{0,5}\}$, then $s^{**} = \{u^1_{-1,1}, u^1_{-1,2}, u^1_{-1,3}, u^1_{-1,4}, u^1_{-1,5}, u^1_{-1,6}\}$, with effective sampling size 6.

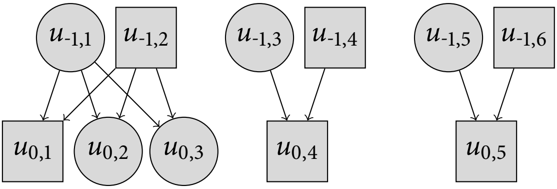


Figure 3.2.: An example of a pedigree graph.

It is convenient to extend the definition of the counting function $N(.)$ so that it counts the number of people included in the sample s^{**} from arbitrary subsets of $A \subseteq \mathcal{U}^p$, counting multiplicities. Any given sample of children s will then induce a *counting measure*, a function defined on all subsets of parents, by:

$$\begin{aligned} N\colon \mathcal{P}(\mathcal{U}^p) &\longrightarrow \mathbb{N} \\ N(A) &\coloneqq \sum_{u \in A} N(\{u\}) \quad \forall A \subseteq \mathcal{U}^p \end{aligned}$$

that associates to a subset of parents A the total number of times they are included in the sample. Here, of course, $N(\{u\}) \coloneqq 0$ if u is not included in the sample of parents s^*.[7]

happens here, I prefer to call the resulting sample a multiset. The reduced set s^* that only enumerates the unique elements of the sample is also called *set-sample* by Särndal et al. (1992: 49).

[7] This justifies the rather clumsy notation $N(\{u\})$ used above.

Finally, the dependence of $N(.)$ on the underlying sample of children and therefore on $\omega \in \Omega^S$ must be made explicit. To do so, one can further extend the definition of $N(.)$ to the two-place function

$$\begin{aligned} N^*: \Omega^S \times \mathcal{P}(\mathcal{U}^p) &\longrightarrow \mathbb{N} \\ N^*(\omega, A) &\coloneqq \sum_{u \in A} |c(u) \cap S(\omega)| \quad \forall A \subseteq \mathcal{U}^p \end{aligned} \tag{3.1}$$

That is, $N^*(\omega, A)$ counts the number of times members of the subset A of the parent generation are included in the sample $S(\omega)$. In this way, $N^*(.,.)$ becomes a *point process* on $\mathcal{U}^p$.

Note that since everyone in the sample of children has both a father and a mother, and since the sample of children is a simple random sample with fixed sample size n, the sample size $n^* \coloneqq |s^{**}|$ (i.e. counting multiplicities) of the parent generation is

$$n^*(\omega) \coloneqq \sum_{u \in \mathcal{U}^p} N^*(\omega, \{u\}) = 2 \sum_{u \in S(\omega)} 1 = 2n$$

for all outcomes of the sampling procedure ω.

The advantages of using the notion of point processes in the context of the induced sampling from the parent generation stem from the ease with which the dependencies between subsets of $\mathcal{U}^p$ can be handled and the possibility to incorporate multisets of observations.[8] In particular, one can define generalisations of the inclusion probabilities introduced in the previous Section. To do so, let

$$\nu(\{u\}) \coloneqq \mathbb{E}(N^*(., \{u\})) = \int_{\Omega^S} N^*(\omega, \{u\})\, d\,\mathrm{Pr}(\omega)$$

where the expectation is taken with respect to the probability distribution of the sampling design. Thus, $\nu(\{u\})$ is the expected number of times that $u \in \mathcal{U}^p$ is included in the sample of parents. It would equal the

[8] A point process is called *simple* if $N^*(., \{u\}) \leq 1$ with probability 1. Thus, we are dealing with non-simple point processes. This must be born in mind since many introductory texts on point processes concentrate on simple processes.

inclusion probabilities $\pi(.)$ if parents would be included at most once into the induced sample. More generally, for subsets $A \subseteq \mathcal{U}^p$ of parents, $\nu(A) \coloneqq \mathbb{E}(N^*(., A))$ is called the *intensity measure* of the point process $N^*(., .)$.

To put the machinery of point processes into action, consider the problem of how to estimate the mean life time of all persons of the parent generation $\mathcal{U}^p$. Whether this is possible is not at all clear at the outset. After all, people with more children will be included with higher probability in the sample and the number of children will normally be connected to survival (or to life time). On the other hand, the sampling theory approach should be agnostic about any such relation between numbers of children and life lengths. In fact, for any parent u, the number of children of u, $|c(u)|$, and her or his life time $T(u)$, are not constrained in any way, not even probabilistically, in the present set-up.

Looking first at the sample total sum of life lengths of parents, one may consider the function

$$\hat{\mathrm{M}}(\omega, T) \coloneqq \sum_{u \in S^*(\omega)} T(u) N^*(\omega, \{u\}) = \sum_{u \in \mathcal{U}^p} T(u) N^*(\omega, \{u\})$$

This is a random function depending on the sampling design. The second expression extends the sum to all parents and makes the handling of $\hat{\mathrm{M}}(., T)$ particularly simple and transparent. The expectation of $\hat{\mathrm{M}}(., T)$ with respect to the design is

$$\mathbb{E}\left(\hat{\mathrm{M}}(., T)\right) = \sum_{u \in \mathcal{U}^p} T(u) \nu(\{u\})$$

because of the additivity of expectations.[9] To compute the intensity measure for singletons, note that

$$\nu(\{u\}) = \mathbb{E}(N^*(., \{u\})) = \sum_{k=0}^{|c(u)|} k \Pr(N^*(., \{u\}) = k)$$

[9] This is a trivial version of *Campbell's* theorem in point process theory (e.g. Reiss 1993: Chap. 5.3) or *Robbin's* formula in the theory of random sets (e.g. Nguyen 2006: Chap. 2). Both are obviously related since samples of the parent generation are random (multi-) subsets of the underlying $\mathcal{U}^p$.

To see how the probability of being exactly k times included in the sample, $\Pr(N^*(.,\{u\}) = k)$, can be computed, one may look back at the example presented at the beginning of the Section. Of the $10 = \binom{5}{3}$ different samples of children, exactly one induces the sample $s^{**} = \{u^3_{\text{-}1,1}, u^3_{\text{-}1,2}\}$ of parents and 3 samples (taking one of the children of the first couple and both the children of the second and third couple) of the form $s^{**} = \{u^1_{\text{-}1,1}, u^1_{\text{-}1,2}, u^1_{\text{-}1,3}, u^1_{\text{-}1,4}, u^1_{\text{-}1,5}, u^1_{\text{-}1,6}\}$. In the remaining 6 cases, the first couple has multiplicities 2 (2 of their 3 children are chosen) and one of the other couples appears once. Thus, $v(\{u_{\text{-}1,1}\}) = v(\{u_{\text{-}1,2}\}) = 1/10 \cdot 3 + 6/10 \cdot 2 + 3/10 \cdot 1 = 1.8$, while $v(\{u_{\text{-}1,j}\}) = 6/10 \cdot 1 = 0.6$ for all other members of the parent generation. Note that indeed $\sum v(\{u\}) = 2 \cdot 1.8 + 4 \cdot 0.6 = 6$, so that the sum of the expected multiplicities gives twice the number of sampled children.

Now consider the slightly more general case where monogamy is no longer assumed, as in the following Figure 3.3. Here, $v(\{u_{\text{-}1,1}\}) = 3/10 \cdot 2 + 6/10 \cdot 1 = 1.2$ but $v(\{u_{\text{-}1,2}\}) = 1/10 \cdot 3 + 6/10 \cdot 2 + 3/10 \cdot 1 = 1.8$ and $v(\{u_{\text{-}1,3}\}) = 3/10 \cdot 2 + 6/10 \cdot 1 = 1.2$ while $v(\{u_{\text{-}1,j}\}) = 0.6$ for the rest of the parents. Once again, $\sum v(\{u\}) = 6$.

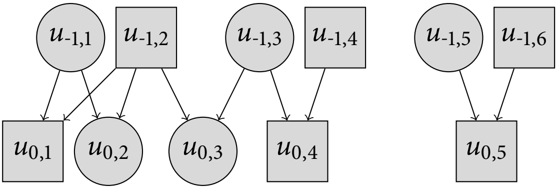


Figure 3.3.: An example of a pedigree graph without monogamy.

The general pattern follows from observing how the multiplicities of a given parent u might arise: She will be included exactly k times in the sample of parents ($k \leq |c(u)|$) if exactly k of her children are included in the sample of children. There are $\binom{|c(u)|}{k}$ ways of selecting k of her children. Setting these aside in sampling the children, one has only to consider the number of samples of size $n - k$ from all the $N_0 - |c(u)|$

children that are not hers. It follows that

$$\Pr(N^*(.,\{u\}) = k) = \binom{|c(u)|}{k}\binom{N_0 - |c(u)|}{n-k} \Big/ \binom{N_0}{n}$$

This is just the hypergeometric distribution. It follows that

$$\begin{aligned} v(\{u\}) &= \frac{n|c(u)|}{N_0} \quad \text{and} \\ \mathbb{E}\left(\hat{\mathrm{M}}(.,T)\right) &= \frac{n}{N_0}\sum_{u\in\mathcal{U}^p} T(u)|c(u)| \end{aligned}$$

This suggests to use the estimator

$$\hat{\mathrm{M}}^*(\omega,T) \coloneqq \frac{N_0}{n}\sum_{u\in\mathcal{U}^p} T(u)N^*(\omega,\{u\}) \,/|c(u)| \tag{3.2}$$

which will be design unbiased for the population total $\sum T(u)$. This is called the *Hansen-Hurwitz* estimator of a total (e.g. Tillé 2006: Chap. 2). It is *not* the better known Horvitz-Thompson estimator which would use weights equal to the inclusion probabilities $\pi(.)$, which in this case would be $\Pr(N^*(.,\{u\}) > 0)$. Since the Horvitz-Thompson estimator correspondingly uses only the set-sample s^*, it can in general not be computed from the sample information available in general surveys: It would require either access to a complete list of labels of the parent generation (and an identification of the label from the respondents) or the possibility to identify siblings among the respondents.

In contrast, despite its apparent dependence on the frame $\mathcal{U}^p$ and on $N^*(.,.)$, neither of which will be known in general, (3.2) *can* be computed from the sample information provided by the children alone. It therefore is a valid statistic for most sampling situations. The only prerequisite is that the number of siblings of the children (the respondents) is known. To see this, note that

$$\hat{\mathrm{M}}^*(\omega,T) = \frac{N_0}{n}\sum_{u\in\mathcal{U}^c} \mathbb{1}[S(\omega)](u)\left(\frac{T(m(u))}{c_m(u)} + \frac{T(f(u))}{c_f(u)}\right)$$

where now $c_m(u)$ and $c_f(u)$ refer to the number of siblings of u (including herself) on her mothers resp. her fathers side. The distinction between maternal and paternal siblings is of course necessary unless strict monogamy is assumed. To illustrate this point, reconsider the last example where the sample of children is $s = \{u_{0,1}, u_{0,2}, u_{0,3}\}$. Then the sample of parents is $s^{**} = \{u^2_{\text{-}1,1}, u^3_{\text{-}1,2}, u^1_{\text{-}1,3}\}$. Thus $\hat{\mathrm{M}}^*$ expressed in terms of the parents becomes

$$\hat{\mathrm{M}}^* = \frac{5}{3}\left(\frac{T(u_{\text{-}1,1}) \cdot 2}{2} + \frac{T(u_{\text{-}1,2}) \cdot 3}{3} + \frac{T(u_{\text{-}1,3}) \cdot 1}{2}\right)$$

Now in terms of the children's sample, the first term arises from both $u_{0,1}$ and $u_{0,2}$ reporting the age of their common mother, so it is reported twice. Furthermore, there are two siblings of this mother. The second term arises similarly. Finally, there is only one report on the age of $u_{\text{-}1,3}$ (from $u_{0,3}$), there is, however, a half-brother of $u_{0,3}$ (through his mother) that must be taken into account and so the denominator is also 2. Note that only the values of $c_m(.)$ and $c_f(.)$ for the respondents are needed in the calculation. On the other hand, neither the labels of the parent generation (the elements of $\mathcal{U}^p$) nor the multiplicities of their inclusion in the sample (the values of $N^*(., \{u\})$) are needed for the calculation of $\hat{\mathrm{M}}^*$.

To get an estimate of the mean life length of the parents, the size of the parent generation $N \coloneqq |\mathcal{U}^p|$ is needed. Now clearly

$$\sum_{u \in \mathcal{U}^c} \frac{1}{c_f(u)} = |\mathcal{U}^m| \quad \text{and} \quad \sum_{u \in \mathcal{U}^c} \frac{1}{c_m(u)} = |\mathcal{U}^f_{\text{-}1}|$$

so that a design unbiased estimator of $|\mathcal{U}^p|$ is

$$\sum_{u \in \mathcal{U}^c} \mathbb{1}[S(\omega)](u) \frac{1}{\pi(u)} \left(\frac{1}{c_f(u)} + \frac{1}{c_m(u)}\right)$$

Once again, for the property of design unbiasedness, it is not necessary to know the actual relatedness of the respondents.

The ratio estimator of the mean life time of the parent generation thus becomes

$$\frac{\sum\limits_{u\in S(\omega)}\left(\frac{T(m(u))}{c_m(u)}+\frac{T(f(u))}{c_f(u)}\right)}{\sum\limits_{u\in S(\omega)}\left(\frac{1}{c_f(u)}+\frac{1}{c_m(u)}\right)} \tag{3.3}$$

where the factor $\pi(.)$ cancels since simple random sampling was assumed. This is no longer design unbiased, but will be approximately so, at least to first order and when both n and N_0 are large (Särndal et al. 1992: Chap. 5.6).

In summary, the classical sampling theory approach provides a design unbiased estimate of the total life time of the parent generation and an approximately design unbiased estimator of the mean life time even though nothing is assumed about the connection between survival times and number of children. But this result should not be overstated: First of all, design unbiasedness is not a panacea, and design unbiased estimators can behave rather erratic. Secondly, the classical sampling theory presupposes at least some knowledge of the underlying labels of the population. This is not in general available in the current problem so that many of the standard estimators (and/or evaluations of their performance) are unavailable. And lastly, up until now the estimators constructed from the sampling theory perspective presuppose an extremely simplified and restrictive model world to work at all.

3.4. Assumptions about Independence

One of the restrictions used at the outset in the constructions of the previous two Sections was that instead of the life lengths of the previous generation only the life length of the parent generation was of interest. But such inferences would only rarely be useful. To extend the results to all members of the previous generation, one may assume that mortality is in a reasonable sense 'independent' of whether or not someone of the previous generation had children. Such an assumption would open the possibility to draw at least partial inferences about the mortality of the whole previous generation.[10]

But how can the notion of 'independence' be made precise so that it can serve its purpose? What is meant when it is assumed that two variables are independent? There are basically two different approaches to an explication. A first one conceptually refers to the process that brings about the facts recorded by statistical variables. This might be what most statisticians and econometricians have in mind when they speak of a *data generating process*. There is, however, some ambiguity in this expression because one needs to distinguish between processes which create data and processes which bring about the facts that, subsequently, can be recorded as data. I propose to use the expression 'data generating process' only in the former sense, for example, when referring to a survey, and speak of "social processes" when referring to processes that one can sensibly think of as creating facts.

A second approach avoids speculations about social processes and concentrates instead on the recorded facts. The approach follows closely the survey sampling tradition where one considers a two-dimensional variable

$$(X, Y) : \mathcal{U} \longrightarrow \mathcal{X} \times \mathcal{Y} \tag{3.4}$$

In this context, the presumption is that (X, Y) records the facts prevailing in the population $\mathcal{U}$ according to some categorisation expressed by $\mathcal{X} \times \mathcal{Y}$. This is a particular way to record the facts, a way which I have called a *statistical variable*. The approach avoids to say anything about how facts are brought about. It only stipulates that the facts are reasonably recorded by using $\mathcal{X} \times \mathcal{Y}$ and that the facts can be ascertained for any person in the population $\mathcal{U}$.

The first approach has to add further structure to (3.4) in order to express 'independence'. It must express the idea that values of X and Y result from some processes and then assume that these processes develop

[10] The inferences will still be partial since it will not be possible to estimate survival probabilities for children and young adults below the minimum child bearing age.

independently. In other words, a model for the social processes that create facts must be proposed. In addition to the assumption that the facts are reasonably described by $\mathcal{X} \times \mathcal{Y}$, one needs a particular way to model the emergence of these facts. If a probabilistic model—as in the "simple device"—is adopted, then the emergence of values can be described by a probability space and a reasonable explication of independence becomes *stochastic independence* between random variables defined on that probability space.

The mathematical definition of stochastic independence requires (after the introduction of a suitable probability space) that all joint probabilities are equal to the product of the respective marginal probabilities. This might be done by writing

$$(X, Y)\colon \mathcal{U} \times \Omega \longrightarrow \mathcal{X} \times \mathcal{Y} \tag{3.5}$$

where Ω is equipped with a set of subsets of Ω, say $\mathcal{B}$, for which a probability Pr(.) is defined. Then a probability distribution for (X, Y) can be introduced by setting

$$\Pr((X, Y)(u, .) \in A) \coloneqq \Pr(\{\omega \in \Omega \mid (X, Y)(u, \omega) \in A\})$$

provided that $\{\omega \in \Omega \mid (X, Y)(u, \omega) \in A\}$ is an element of $\mathcal{B}$.

For each member u of the population, stochastic independence can now be defined by

$$\forall A \subseteq \mathcal{X}\ \forall B \subseteq \mathcal{Y}:\ \Pr(X(u, .) \in A \mid Y(u, .) \in B) = \Pr(X(u, .) \in A) \tag{3.6}$$

where both $X(u, .)^{-1}(A)$ and $Y(u, .)^{-1}(B)$ are elements of $\mathcal{B}$ so that probabilities for these 'events' can be deduced. Note that stochastic independence can be defined for each member of the population $\mathcal{U}$. And it is an extra assumption that independence holds for all $u \in \mathcal{U}$. It must be stressed that independence of $(X, Y)(u, .)$ from $(X, Y)(u', .)$, while defined in parallel to the definition (3.6) in probabilistic approaches, is used for purposes quite different from the definition of the independence of X from Y. The former is a constitutive decision about the model structure and must be distinguished from assumptions like independence of $X(u,.)$ from $Y(u,.)$ for all $u \in \mathcal{U}$.[11]

Mathematicians and statisticians often call elements of $\mathcal{B}$ 'events' even though, of course, nothing has to happen for such 'events' to be defined. Rather, the notion of 'events' described by a system of subsets $\mathcal{B}$ is a static notion quite far removed from dynamic concepts. While such parlance can only present stylised facts and not their bringing about, the mathematical definition of independence can be turned into a procedural explication by referring to random generators that work physically independent from each other. For example, one might create values for X with a die, and values for Y with another die, and assume that both processes work independently from each other. The procedural explication of stochastic independence refers to causally unrelated methods of producing real events that can be described by a probability model.

This approach presupposes a probabilistic model for the social processes that create facts. While the introduction of a probability space might seem to be an extraneous and artificial addition to the description of the emergence of social facts, the obvious advantage of the approach is that it can rely on the calculus of probability, including the many results known about independent variables.

On the other hand, the approach may be judged to be far to restrictive, presupposing that probability models accurately describe the constitution as well as the changes of statistical variables recording social facts. An alternative approach should therefore refer only to the eventually realised data, i.e. the data generating process in the restricted sense. Turning back to the formulation (3.4), a description of social facts, one may try to define independence not with respect to some probability space but with respect to the relative frequencies that describe social facts. To this

[11] While the condition (3.6) seems innocuous and is often treated in textbooks as a basic definition that does not deserve much comment, the mathematical strength of the condition can not be overemphasised. E.g., Holbrook (1981) shows that if $(X(\omega), Y(\omega))$ are independent uniform random variables defined on $\Omega = [0,1]$ with Lebesgue measure, then $(X, Y)\colon \Omega \longrightarrow [0,1] \times [0,1]$ is a space-filling curve. Further examples from many areas of mathematics are discussed by Kac (1959).

end, write $\mathrm{P}(X \in A)$ for the relative frequency, i.e.

$$\begin{aligned} \mathrm{P}\colon \mathcal{P}(\mathcal{X}) &\longrightarrow [0,1] \\ \mathrm{P}(X \in A) &\coloneqq \left|\{u \in \mathcal{U} \mid X(u) \in A\}\right| / |\mathcal{U}| \quad \forall A \in \mathcal{P}(\mathcal{X}) \end{aligned} \tag{3.7}$$

This is, of course, a probability measure, albeit a special one. There is no additional probability space Ω needed for its construction. Furthermore, it is defined not for each member of $\mathcal{U}$ but only for the population $\mathcal{U}$ as a whole. It simply counts subsets of the population $\mathcal{U}$ and relies only on the statistical variables introduced in (3.4).

The intuition behind this approach, rather different from the probabilistic version, is that X is 'independent' from Y if the distribution of X does not depend on subsets of $\mathcal{U}$ that can be selected by considering values of Y alone. The advantage of such a definition is that an explicit reference to the processes creating the values of X and Y would be superfluous.

One may then try to copy the definition of stochastic independence from the probabilistic version (3.6) by requiring

$$\forall A \subseteq \mathcal{X} \; \forall B \subseteq \mathcal{Y}: \; \mathrm{P}(X \in A \mid Y \in B) = \mathrm{P}(X \in A) \tag{3.8}$$

There is, however, a problem resulting from the fact that the equality formulated in (3.8) can not, in general, be satisfied. A small example will show this. Assume that $\mathcal{X} \coloneqq \mathcal{Y} \coloneqq \{0,1\}$, $\mathcal{U}$ has 10 members, and the marginal distributions are as follows:

	$X = 0$	$X = 1$	
$Y = 0$	n_{11}	n_{12}	4
$Y = 1$	n_{21}	n_{22}	6
	3	7	10

(3.9)

I now want to make the assumption that X and Y are independent. Following the definition given in (3.8), one would need to find, for example, values of n_{11} and n_{12} such that

$$n_{11} + n_{12} = 4, \quad \frac{n_{11}}{4} = \frac{3}{10}, \quad \text{and} \quad \frac{n_{12}}{4} = \frac{7}{10}$$

But no integral values of n_{11} and n_{12} can be found that satisfy these requirements. The simple consequence is that (3.8) can not, in general, be used to postulate assumptions about independence in finite populations.

In fact, there are four tables with the given marginal frequencies:

$$a)\ \begin{matrix}0 & 4\\ 3 & 3\end{matrix}\quad b)\ \begin{matrix}1 & 3\\ 2 & 4\end{matrix}\quad c)\ \begin{matrix}2 & 2\\ 1 & 5\end{matrix}\quad d)\ \begin{matrix}3 & 1\\ 0 & 6\end{matrix} \tag{3.10}$$

The corresponding relative frequencies of these tables can be embedded in the simplex $\{(p_1, p_2, p_3, p_4) \mid p_i \geq 0, \sum p_i = 1\}$. In Figure 3.4, the four points are shown together with the surface of all points that satisfy the independence formulation 3.8. It is plain that it will be the exception rather than the rule to find a point on the surface of independence for any given set of marginal frequencies.

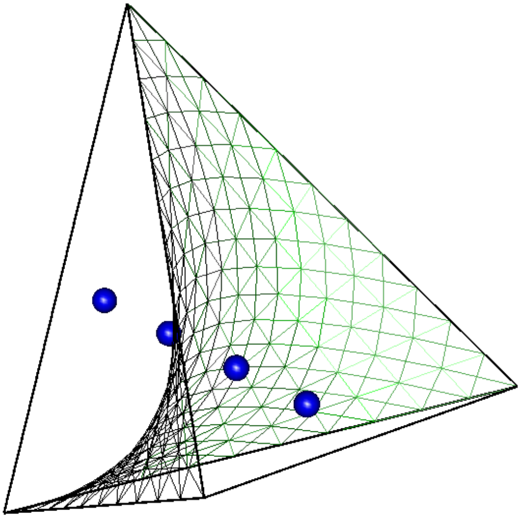


Figure 3.4.: The simplex of probabilities in 2 × 2 tables, the surface of independence, and the 4 tables with margins given by Example 3.9.

If one nevertheless wants to follow this second approach one needs a weaker formulation that allows for approximate equalities between conditional distributions. I will use the following definition:

X is Δ-independent from Y w.r.t. the partition $\mathcal{Y}_p$ of $\mathcal{Y}$ if

$$\max\{|\mathrm{P}(X \in A \mid Y \in B) - \mathrm{P}(X \in A)| \mid A \subseteq \mathcal{X}, B \in \mathcal{Y}_p\} \leq \Delta \tag{3.11}$$

The idea behind this definition is that conditional distributions of X should be approximately equal among the subsets of $\mathcal{U}$ induced by a given partition of $\mathcal{Y}$. The partition $\mathcal{Y}_p$ (instead of the singletons of $\mathcal{Y}$) figures in the definition since it will often only be necessary to distinguish less detail than could be achieved with the full information from Y. E.g. in the problem treated here, $Y(u)$ would correspond to the names (and numbers) of children of u. But what is needed, at least in this Section, is only the information whether u had children or not ($|c(u)| > 0$ or $|c(u)| = 0$).

The earlier formulation in (3.8) can then be seen as a limiting case where $\Delta = 0$ and $\mathcal{Y}$ is partitioned into its one-element subsets. But, as the example has shown, these strong requirements are rarely fulfilled. On the other hand, given any partition of $\mathcal{Y}$, one can always find a minimal Δ such that X is Δ-independent from Y. To illustrate with the example, one finds for the table b) in (3.10):

$$\begin{aligned}
|\mathrm{P}(X = 0 \mid Y = 0) - \mathrm{P}(X = 0)| &= |1/4 - 3/10| = 0.050 \\
|\mathrm{P}(X = 0 \mid Y = 1) - \mathrm{P}(X = 0)| &= |2/6 - 3/10| = 0.033 \\
|\mathrm{P}(X = 1 \mid Y = 0) - \mathrm{P}(X = 1)| &= |3/4 - 7/10| = 0.050 \\
|\mathrm{P}(X = 1 \mid Y = 1) - \mathrm{P}(X = 1)| &= |4/6 - 7/10| = 0.033
\end{aligned}$$

The minimal Δ is therefore 0.05 w.r.t. the partition of $\mathcal{Y}$ into the sets $\{0\}$ and $\{1\}$.

Note the asymmetry of the definition: While X is 0.05-independent from Y, the minimum of Δ for Y given X follows from

$$\begin{aligned}
|\mathrm{P}(Y = 0 \mid X = 0) - \mathrm{P}(Y = 0)| &= |1/3 - 4/10| = 0.0667 \\
|\mathrm{P}(Y = 0 \mid X = 1) - \mathrm{P}(Y = 0)| &= |3/7 - 4/10| = 0.0286 \\
|\mathrm{P}(Y = 1 \mid X = 0) - \mathrm{P}(Y = 1)| &= |2/3 - 6/10| = 0.0667
\end{aligned}$$

$$|\mathrm{P}(Y = 1 \,|\, X = 1) - \mathrm{P}(Y = 1)| \;=\; |4/7 - 6/10| \;=\; 0.0286$$

Thus, Y is only 0.0667-independent from X. The asymmetry may seem unfortunate. But the formulation accurately reflects what is needed in the application: There is information on the relative frequencies of life times of members of the previous generation given that they have children, i.e. $\mathrm{P}(T = . \,|\, |c| > 0)$ is (approximately) known, and this should provide a clue for the frequencies of life lengths of all members of the previous generation, $\mathrm{P}(T = .)$. When it can be stipulated that T is Δ-independent of $\{|c| > 0\}$, then the frequencies are close by definition, $\mathrm{P}(T = . \,|\, |c| > 0) \approx \mathrm{P}(T = .)$. If further $\mathcal{T}$ is finite, then the closeness of relative frequencies implies closeness of conditional means to marginal means and the same holds true for many other functions of the frequencies.

A few immediate implications of the definition are:

a) If X is Δ-independent from Y w.r.t. a partition $\mathcal{Y}_p$, then this is also true for all $\Delta' \geq \Delta$.

b) For a fixed $B \in \mathcal{Y}_p$,

$$\max_{A \subseteq \mathcal{X}} \{|\mathrm{P}(X \in A \,|\, Y \in B) - \mathrm{P}(X \in A)|\} =: \|\mathrm{P}(X \in . \,|\, Y \in B) - \mathrm{P}(X \in .)\|$$

is often called the *total variation distance* between $\mathrm{P}(X \in . \,|\, Y \in B)$ and $\mathrm{P}(X \in .)$. This distance can be expressed in terms of the densities $\mathrm{P}(X = x)$ and $\mathrm{P}(X = x \,|\, Y \in B)$:

$$\max_{A \subseteq \mathcal{X}} \{|\mathrm{P}(X \in A \,|\, Y \in B) - \mathrm{P}(X \in A)|\} = \frac{1}{2} \sum_{x \in \mathcal{X}} |\mathrm{P}(X = x \,|\, Y \in B) - \mathrm{P}(X = x)|$$

This shows that the computation of Δ for a given B can be carried out running only through the singletons of $\mathcal{X}$. There is no need to consider all subsets of $\mathcal{X}$. Only $|\mathcal{X}| \cdot |\mathcal{Y}_p|$ absolute differences must be computed.[12]

[12] Strasser (1985: 5–7) proofs this and similar relations in a rather general setting.

c) If X is Δ-independent from Y w.r.t. a partition $\mathcal{Y}_p$, and B_1 and B_2 are both elements of $\mathcal{Y}_p$, then

$$\max\{|\mathrm{P}(X \in A \mid Y \in B_1 \cup B_2) - \mathrm{P}(X \in A)| \mid A \subseteq \mathcal{X}\} \;\leq\; \Delta$$

This follows since if $B_1 \neq B_2$, then $B_1 \cap B_2 = \emptyset$ ($\mathcal{Y}_p$ is a partition) so that

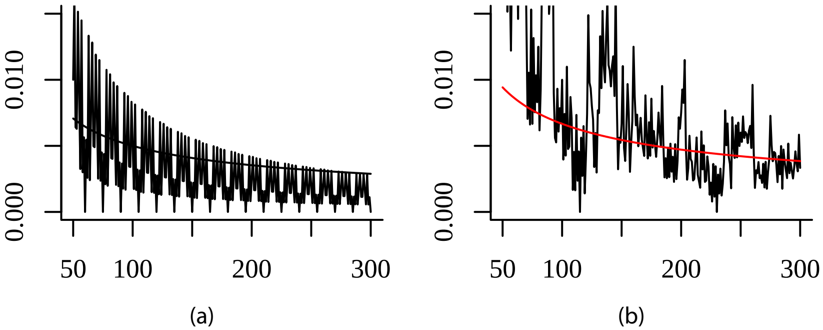
$$\begin{aligned}\mathrm{P}(X \in A \mid Y \in B_1 \cup B_2) &= \\ \mathrm{P}(X \in A \mid Y \in B_1)&\frac{\mathrm{P}(Y \in B_1)}{\mathrm{P}(Y \in B_1) + \mathrm{P}(Y \in B_2)} \\ + \mathrm{P}(X \in A \mid Y \in B_2)&\frac{\mathrm{P}(Y \in B_2)}{\mathrm{P}(Y \in B_1) + \mathrm{P}(Y \in B_2)}\end{aligned}$$

d) As an implication of (b) one finds: If X is Δ-independent from Y w.r.t. a partition $\mathcal{Y}_p$, then X is also Δ-independent from Y w.r.t. any coarser partition of $\mathcal{Y}$.[13] In particular, if X is Δ-independent from Y w.r.t. to the partition of $\mathcal{Y}$ into one-element subsets, it is Δ-independent from Y w.r.t. to any partition of $\mathcal{Y}$.

I therefore simply say that X is Δ-independent from Y if this is true w.r.t. the partition of $\mathcal{Y}$ into one-element subsets.

As shown by the example of the 2×2 table, one can not expect a small Δ even when the data are as close as possible to stochastic independent. It is natural to ask for the smallest possible Δ achievable in a finite population of size n and a given set of conditional frequencies $\mathrm{P}(X = x \mid Y = y)$. I.e. given the distribution of the life lengths of the parent generation and the total size n of the previous generation, what is the smallest achievable value of Δ? To facilitate the discussion, suppose that $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ so that we deal with a 2×2 table. Suppose further that both $\mathrm{P}(X = 0 \mid Y = 1)$ and $\mathrm{P}(Y = 1)$ are fixed at some rational numbers and that n is given. Now choose natural numbers closest to $n\mathrm{P}(Y = 1)$ and $n\mathrm{P}(X = 0, Y = 1)$ and choose $n\mathrm{P}(X = 0, Y = 0)$ in such a way that Δ is minimised. It might be

[13] A partition $\mathcal{Y}'_p$ is *coarser* than a partition $\mathcal{Y}_p$, denoted by $\mathcal{Y}_p \subseteq_p \mathcal{Y}'_p$, if for each $B \in \mathcal{Y}_p$ there is an element $B' \in \mathcal{Y}'_p$ such that $B \subseteq B'$.
As a consequence, one can extend Δ-independence w.r.t. $\mathcal{Y}_p$ to Δ-independence w.r.t. the σ-algebra generated by the partition $\mathcal{Y}_p$.



(a)

(b)

Figure 3.5.: The minimal Δ achievable. a) $\mathrm{P}(Y=1) = 6/10$, $\mathrm{P}(X=0 \mid Y=1) = 1/3$ corresponding to 3.10b). b) $\mathcal{X} = \{0,1,2,3\}$, $\mathrm{P}(Y=1) = 313/521$, $\mathrm{P}(X=0 \mid Y=1) = 14/313$, $\mathrm{P}(X=1 \mid Y=1) = 41/313$, $\mathrm{P}(X=3 \mid Y=1) = 149/313$. The smooth curve is proportional to $1/\sqrt{n}$, and thus also proportional to the standard errors computed either for a simple random sample or for n independent realisations of respective random variables.

conjectured that Δ converges rapidly with n to 0. But this is not true, the convergence is far from uniform, and it is not as fast as is often suggested by statistical reasoning. Figure 3.5 shows the minimal Δ for n from 50 to 300 together with an indication of the order of the estimated standard error. Also indicated is a case with $|\mathcal{X}| = 4$, where the behaviour is even less regular.

A further example may help to illustrate the conceptual difference between Δ-independence and stochastic independence. Values of the variable (X, Y) are created with a pair of dice such that the dice operate independently. Throwing the pair of dice 60 times, again independently, the following values might arise:

	$Y=1$	$Y=2$	$Y=3$	$Y=4$	$Y=5$	$Y=6$	
$X=1$	4	3	0	1	2	1	11
$X=2$	3	1	0	3	0	2	9
$X=3$	3	0	0	3	3	1	10
$X=4$	2	1	4	2	0	1	10

	$Y = 1$	$Y = 2$	$Y = 3$	$Y = 4$	$Y = 5$	$Y = 6$	
$X = 5$	3	1	0	3	2	0	9
$X = 6$	2	4	1	1	2	1	11
	17	10	5	13	9	6	60

Now, from a procedural point of view, the processes creating the values of X and Y are clearly independent, and this provides one possible explication of the independence of the two variables: that their values are generated by independent processes. Note that this explication completely ignores the actually realised values. On the other hand, one might as well ask for the degree of independence exhibited by the realised data. Given a partition of $\mathcal{Y}$, I can ask to which degree X is Δ-independent from Y w.r.t. this partition. Using the partition $\mathcal{Y}_p \coloneqq \{\{1\}, \ldots, \{6\}\}$, the task is to find the smallest Δ such that X is Δ-independent from Y w.r.t. this partition. In a first step, one may consider the restricted maximum

$$\max_{x \in \mathcal{X}, y \in \mathcal{Y}} \{|\mathrm{P}(X = x \mid Y = y) - \mathrm{P}(X = x)|\}$$

that only takes into account one-element subsets of $\mathcal{Y}$. This is easy to calculate and yields 0.633 (arising from the entry $X = 4$, $Y = 3$). Such a large value is hardly useful and the conditional density of X changes heavily with the value of y chosen.

To compute the final value of Δ, one has to compute the maximum with respect to all subsets of $\mathcal{X}$. By property b) above, it is enough to compute the sum of the absolute differences between the conditional and marginal densities for all fixed y. This gives 0.115, 0.333, 0.650, 0.226, 0.316, 0.183. Thus, the smallest value of Δ achievable is 0.65. The conditional distribution behaves only slightly worse than the conditional density.

In summary, Δ-independence of one variable from another is conceptually different from stochastic independence. Statistical variables with small Δ-independence obviously need not be created from stochastically independent random variables. In fact, Δ-independence is defined without reference to an additionally introduced probability space Ω. It only refers to the recorded data.

Nor does stochastic independence imply Δ-independence for a reasonably small Δ. This is clear from the dice example. It is only within the model world, and only when further assumption are made, that Δ-independence nearly follows from stochastic independence, using a form of the strong law of large numbers. *If* data are generated by an appropriate random device, *if* draws from the device are stochastically independent, *if* both $\mathcal{X}$ and $\mathcal{Y}_p$ are finite, *if* the number of draws is potentially infinite, and, of course, *if* $X \perp\!\!\!\perp Y$, then the event that

$$\max_{y_p \in \mathcal{Y}_p} \|\mathrm{P}_n(X \in . \mid Y \in y_p) - \mathrm{P}_n(X \in .)\| > \Delta$$

will have probability 0 for all $\Delta > 0$ and n large enough. Here, $\|.\|$ is the total variation distance and $\mathrm{P}_n(.)$ is the empirical distribution of the first n elements of a sequence $(X, Y)_1, (X, Y)_2, \ldots, (X, Y)_n, \ldots$ of realisations of (X, Y). But this result has no operational meaning: The conclusion can only be formulated within the probability model.

On the other hand, the concept of Δ-independence can be used easily to formulate assumptions about the distribution of life lengths of the previous generation based on knowledge of the distribution of the life lengths of the parent generation. If it can be argued that the distribution of life lengths T is Δ-independent from $\{|c(.)| > 0\}$ for a small enough Δ, then knowledge of the life lengths of the parent generation implies knowledge of the distribution of life lengths of the previous generation up to an error of at most Δ.

A claim of Δ-independence is a claim pertaining to the facts recorded by the variables under consideration. Therefore, arguments in favour of the claim—and its criticism—must rest on references to other pertinent facts, e.g. information on life lengths of the generation from other sources.[14] Arguments referring to genetic inheritance, living conditions, and historical circumstances are irrelevant in the discussion of such claims. They can not even be incorporated into the formulation of Δ-independence.

[14] It should be clear, however, that the arguments are constrained by the framework introduced in Section 3.1.

3.5. The "Simple Device"

In contrast to the approach based on statistical variables, the "simple device"—when taken as the general advice to express everything in terms of a probability model—would require to account for the emergence of all the facts including the development of the ancestral graph. This is sometimes attempted in the biological sciences (e.g. Whittemore/Halpern 1994), but it would be certainly too much to require such a detailed model only to answer a simple question on the distribution of life lengths.

In fact, the "simple device" can be made to work within a much more limited scope. Considering only the previous generation (ignoring the sampling from the children's generation), one may construct a family of random variables

$$(T, C)\colon \mathcal{U} \times \Omega \longrightarrow \mathcal{T} \times \{0, 1, \ldots, K\}$$

where $(T, C)(u, .)$ represents life length and number of children of the element $u \in \mathcal{U}$, respectively. The probability space $(\Omega, \mathcal{B}, \Pr)$ and the random variable (T, C) is such that there is at least one ω such that $(T, C)(u, \omega)$ equals the life length and number of children for all $u \in \mathcal{U}$. Suppose further that the random variables constructed for different members u of the population are stochastically independent and identically distributed (a constitutive model assumption). Following the notation developed in Chapter 2, the coarsened variables (representing information on the parents only) can be written as

$$\begin{aligned}(T, C)^*\colon \mathcal{U}\times\Omega \longrightarrow \{\{(0, 1)\}, \{(1, 1)\}, \ldots, \{(\tau, K)\}\}\cup\{\mathcal{T}\times\{0\}\} \\ \subset \mathcal{P}(\mathcal{T} \times \{0, \ldots, K\})\end{aligned}$$

i.e. $(T, C)(u, .)$ is observed exactly if and only if $C(u, .) > 0$ (u has at least one child). Otherwise, any value of $\mathcal{T}$ might have happened. The implied probabilities for (T, C) can conveniently be arranged in a table

$\Pr(T = 0, C > 0)$	$\ldots$	$\Pr(T = \tau, C > 0)$	$\Pr(C > 0)$
$*$	$\ldots$	$*$	$\Pr(C = 0)$
$*$	$\ldots$	$*$	

where the $*$ indicate unknown probabilities and the values of $C(u,.)$ are coarsened to the set $\{0,1\}$ with 1 indicating $C(u,.) > 0$.

As introduced up to now, the “simple device”, by expressing the situation in terms of random variables, forces the assumption that the probability of having children is known. This follows since random variables are functions on the known domain $\mathcal{U} \times \Omega$ so that both the number of elements of $\mathcal{U}$ and the probability of $\{C(u,.) > 0\}$ is supposed to be known. Even the values of $C(u,\omega)$ must be supposed known for all $u \in \mathcal{U}$ and the realised ω. The assumption contradicts the original description of the problem where the size of the previous generation is unknown. Of course, the model may be amended by restricting attention to the subset of $\mathcal{U}$ such that $C(u,.) > 0$. In probabilistic terms, one needs to consider the conditional distribution of $(T,C)(u,.)$ given the event $\{C(u,.) > 0\}$. The existence of (and reference to) elements of childless members of the previous generation then becomes part of the probability model itself, further separating the “simple device” from its sampling theory counterpart. However, I will discuss the conditional version of the model only in the next Section.

In order to facilitate the comparison with results from Chapter 2 and to ease the notational burden, I will for the rest of the Section drop the reference to individuals u so that (T,C) and $(T,C)^*$ are just pairs of random variables defined on a common probability space. Within the present model, the set of underlying distributions of (T,C) compatible with a given distribution of $(T,C)^*$ is easily described: it is the simplex $\{\Pr(T=t, C=0) \mid t \in \{0,\dots,\tau\}\}$. It is also easy to see that there is a unique CAR solution: Remember from 2.21 that the joined distribution of (Y, Y^*) satisfies CAR if

$$\Pr(Y = y \mid Y^* = y^*) = \Pr(Y = y \mid Y \in y^*)$$

In the present situation, $C(\omega) = 0$ (no children) is equivalent to $T^*(\omega) = \mathcal{T}$, so that $\{C = 0\} := \{\omega \in \Omega \mid C(\omega) = 0\} = \{\omega \mid T^*(\omega) = \mathcal{T}\}$. Now

$$\begin{aligned}
&\Pr(T = t, C = 0 \mid (T,C)^* = \mathcal{T} \times \{0\}) \\
&= \Pr(T = t \mid T^* = \mathcal{T}, C = 0) \\
&= \Pr(T = t \mid C = 0)
\end{aligned}$$

and

$$\Pr(T = t \mid C = 0) = \Pr(T = t) = \Pr(T = t \mid T \in \mathcal{T})$$

if and only if T and C are stochastically independent ($T \perp\!\!\!\perp C$). It follows that $((T, C), (T, C)^*)$ is CAR if and only if $T \perp\!\!\!\perp C$. Since the conditional distribution of T given $\{C > 0\}$ is known, the CAR condition uniquely identifies the joined distribution of (T, C) from the known (estimated, approximated, assumed, ...) distribution of $(T, C)^*$.[15]

Within the present rather simplistic probabilistic model, stochastic independence or, equivalently, the CAR condition takes the role of Δ-independence in the approach based on relative frequencies. But what exactly is the role of the probability model? The model as presented does not pretend to refer to the process by which life lengths are determined. It does not even provide the means to express pedigree, or inheritance, or when people do have children. And certainly life lengths are not created by the roll of a die. An argument in favour of such a model must necessarily rely on an "as if" reasoning: that the distribution of life lengths may be treated for the current purpose "as if" generated by a random device. As argued in section 2.2.1, the adoption of a particular probabilistic model is best seen as decision taken by the researcher. Arguments in support of the decision for using a probabilistic model can not simply refer to reality or data. But what kind of arguments can provide justifications for the decision?

Jerzy Neyman (1960: 629) explicitly argued for the use of stochastic models acknowledging their "as if" character:

> The essence of dynamic indeterminism in science consists in an effort to invent a hypothetical chance mechanism, called a "stochastic model", operating on various clearly defined hypothetical entities, such that the resulting frequencies of the various possible outcomes correspond approximately to those actually observed.

[15] But note that this joined distribution can not, in general, coincide with any frequency distribution realised using the same model (cf. Example 3.9).

He also emphasises that the term 'indeterminism' relates "not to the phenomena themselves but to our approach to these phenomena" (Neyman 1960: 626).

He then goes on to formulate criteria for the 'scientific value' of such models: 'broad applicability' and 'identifiability'. He introduces these criteria only by example. But even a cursory reading shows that the model introduced here is neither broadly applicable nor identifiable in his sense. In particular, the "simple device", expressing incomplete observations in a particular way, is the only aspect of 'broad applicability' of the model. However, this is a methodological recipe, a mere template. As such it has no 'scientific value'. It must be amended within a given context. But then its merits must be judged separately in each application. This can be done in the present case. But it can only be done within a much more elaborate model where it is possible to argue about the details of the model based both on consequences of the model and the known facts. It is only in relatively rich models that the strengths of the "simple device", the possibility to discuss in a principled way consequences of different assumptions of missingness, can be achieved. The "simple device" per se will not necessarily provide such a framework.

To be more precise, the set of distributions of (T, C) compatible with $(T, C)^*$ is the simplex

$$\{\Pr((T,C) = (t,0) \mid t \in \mathcal{T}, \\ \sum_{t=0}^{\tau} \Pr(T = t, C = 0) = \Pr((T,C)^* = \mathcal{T} \times \{0\}\}$$

In the absence of further model details, the CAR solution is but a particular point in this set. There is no way to argue persuasively for a subset of compatible distributions, let alone for one particular solution, except, of course, that the CAR condition would justify a particularly simple statistical treatment. But simplicity of statistical treatment can not be an argument in judging facts.

Furthermore, if the probability $\Pr((T,C)^* = \mathcal{T} \times \{0\})$ is set equal to the observed frequency $\mathrm{P}(|c(.)| = 0)$, then the set of compatible distributions of the "simple device" coincides with the set-valued sample based solution favoured by Manski (2003) and others. As argued in the previous Chapter, the equality is reassuring from a conceptual point of view: The probability model does not introduce additional aspects that prejudice the solution of the incomplete data problem. But many who think of the "simple device" as a mere statistical "method" for solving their problem will be dissatisfied with this equality since it generally generates rather large intervals of possible values.[16]

But the problem turns out to be worse. As witnessed by the dice example of the previous Section, the connection between an assumed probability model and (realised) values is rather weak, and in general, the connection is even weaker when the model is confronted with observations: Even if observations are generated from a given probability model with probability measure Pr(.) and assuming that n observations are independently generated from this measure, then there is no function of the observations (i.e. a statistic) such that the maximum deviation of empirical frequencies (or kernel estimators etc.) from their theoretical counterparts converges to 0 in the variation distance with the number of observations.[17] The result applies in particular to the "simple device". No amount of observations will allow to narrow the gap between data and model. Neither the rather broad set of compatible distributions nor the CAR condition can be justified by reference to the observations alone. Therefore, arguments for adopting the CAR condition must be sought based on the probability model itself. This can be done, but it requires an elaborated model that goes beyond the mere introduction of some random variables. Note the contrast to the sampling perspective: there, the set of compatible frequencies can be directly compared to frequencies derived either from further data or from other sources. Accordingly, both, arguments for the assumption of Δ-independence and its criticism

[16] E.g., in the discussion of Manski and Horowitz (2000: 86), Raghhunathan complained: "I am afraid that I agree with Cochran (1977) that such an approach is so conservative as to be of little value in most practical settings for inferential purposes."

[17] This might appear to be a somewhat trivial result since empirical distributions always have maximal variation distance from an absolutely continuous probability measure. The negative result is, however, not restricted to the divide between discrete and absolutely continuous distributions. See Section 2.2.1 for general comments. Devroye/Győrfi (1990) present further details on the general negative result.

can be assessed based on empirical information alone.

There is yet another problem with the "simple device" in the present case: As I argued in Section 3.3, sampling the children induces the possibility of including a parent several times without this being known to the surveyor. Even within a probabilistic model, one strategy to deal with the issue is to keep with the classical sampling approach. One would treat the sampling of the children (and the induced sampling of the parents) as a process independent of the underlying probability space of the "simple device". This is what is regularly done in the model assisted approach to sampling.[18] But since the argument in Section 3.3 includes a reference to the number of children, and since values of statistical variables are fixed numbers, one would be induced to treat that number as fixed as well in the probabilistic model of the "simple device". This in turn would require a reformulation of the CAR condition, and of stochastic independence, where the number of children may now only appear on the right side of the conditioning bar.

One might, however, try to follow the opposite strategy and formulate the problem completely within a probabilistic framework. After all, not knowing the number of times a particular sample member is included in the sample might be seen as a further level of incompleteness. But this approach turns out to be difficult at best, without offering an appreciable advantage. In fact, in trying to do so one has to model the ancestral graph and its development. In consequence, one can not take the reference to individuals (the sets $\mathcal{U}_t$ and $\mathcal{U}_{t+1}$) as fixed. Their coming about must become part of the model. One can then condition on the realised units of the children's generation and look backwards in time to produce realisations of the complete previous generation (not only the parents). This would provide a model both for the multiplicities in a sample and the number (and life lengths) of childless people. But the childless people of the model are now just creations of the model without an identifiable counterpart in (partially) recorded data.[19] It is at least difficult to see

[18] See e.g. Särndal et al. (1992). For a more comprehensive discussion of a joint sampling and probabilistic approach see Rubin-Bleuer/Schiopu Kratina (2005) and Hansen et al. (1983).

[19] McCullagh (2008) discusses problems arising from randomly labelled units in the context of binary random effects models. In his his article of 2005 he briefly refers to

how even summary statistics about these phantoms could be related to known facts. Even if this problem is dismissed in favour of a purely model based interpretation, the problem of multiple inclusion into the sample still has no obvious solution. In particular, the analysis of the weighted solution (3.3) appears to be extremely difficult.

3.6. Left Truncated Data

In order to arrive at a more realistic model, one further aspect has to be incorporated into the description of the situation. Up to now, no restrictions between life lengths and number of children were considered. But children and young adults below the minimum child bearing age can not have children. Therefore, no deaths before the minimum child bearing age can be recorded in the parent sample and nothing can be said about the distribution of very short life lengths.[20]

To discuss the issue, consider a slight extension of the previous model where the statistical variables (T, C) are defined as

$$(T, C)\colon \mathcal{U} \longrightarrow \mathcal{T} \times \big(\mathcal{T} \cup \{-1\}\big) \tag{3.12}$$

As before, $T(u)$ denotes u's life length for each $u \in \mathcal{U}$. In contrast to the previous Section, however, $C(u)$ now records the age at which u became for the first time mother (or father) of a child. If u stays childless throughout his or her lifetime, $C(u)$ is set to -1. To facilitate the discussion, I will subsequently only deal with women (and still denote the population by $\mathcal{U}$, the population of mothers by $\mathcal{U}^p$ etc.).[21]

what he calls "block factors", i.e. unlabelled factors in a regression that just indicate membership in an otherwise uncharacterised group. He shows that in both cases unlabelled units do create problems of reference and interpretation. The present case may be considered as even worse since the labels only exist within the model, while in McCullagh's cases they do have a real counterpart.

[20] I still assume that information from all children is available, ignoring their mortality up to the interview date.

[21] The same reasoning, however, applies to men with obvious minor modifications.

The available data still refer to the subset of women that had at least one child, say

$$\mathcal{U}^p \;:=\; \{u \in \mathcal{U} \mid C(u) \geq 0\}$$

But with the introduction of the variable $C(.)$ (the age at which a first child was born) it is now possible to discuss the impact of a lower truncation point of observed life times on the assessment of the distribution of life lengths.

Following the basic idea of the Kaplan-Meier procedure, one may use rates to assess the distribution of T. Assuming complete observations it would be possible to define both a *risk set*

$$\mathcal{U}_t \;:=\; \{u \in \mathcal{U} \mid T(u) \geq t\}$$

i.e. the set of all members of $\mathcal{U}$ who are alive just before age t, and an *event set*

$$\{u \in \mathcal{U}_t \mid T(u) = t\}$$

consisting of those members of $\mathcal{U}_t$ who actually died at age t.[22] From these sets one can calculate rates

$$r(t) \;:=\; \frac{|\{u \in \mathcal{U}_t \mid T(u) = t\}|}{|\mathcal{U}_t|} \;=\; \mathrm{P}(T = t \mid T \geq t)$$

which can be used to find the survivor function

$$\mathrm{P}(T \geq t) =: G(t) \;=\; \prod_{j=0}^{t-1} (1 - r(j))$$

But since the data only refer to $\mathcal{U}^p$, neither the cardinality of the risk set nor that of the event set is known and consequently one cannot calculate the rates $r(t)$. One can only try to estimate these rates, but this will then require an assumption. The assumption will roughly be that mortality

[22] I am once again slightly misusing notation by reusing the symbol $\mathcal{U}_t$ to denote not the (previous) generation but a subset of $\mathcal{U}$, the set at risk of dying.

does not depend on whether and when people became mothers and fathers. More precisely, the assumption is

$$\begin{aligned}\tilde{r}^*(t) &:= \mathrm{P}(T = t \mid T \geq t, 0 \leq C \leq t) = \frac{|\{u \in \mathcal{U}_t^p \mid T(u) = t\}|}{|\mathcal{U}_t^p|} \\ &\approx \frac{|\{u \in \mathcal{U}_t \mid T(u) = t\}|}{|\mathcal{U}_t|} = \mathrm{P}(T = t \mid T \geq t) = r(t)\end{aligned}$$

where the risk set on the left-hand side is now defined by

$$\begin{aligned}\mathcal{U}_t^p &:= \{u \in \mathcal{U}^p \mid T(u) \geq t, 0 \leq C(u) \leq t\} \\ &= \{u \in \mathcal{U} \mid T(u) \geq t, 0 \leq C(u) \leq t\}\end{aligned}$$

Looking at the expression of the rates in terms of conditional frequencies

$$\mathrm{P}(T = t \mid T \geq t, 0 \leq C \leq t) \approx \mathrm{P}(T = t \mid T \geq t)$$

it becomes apparent that this is just a dynamic version of Δ-independence introduced in Section 3.4. In other words, $\{T = t\}$ is Δ-independent from $\{0 \leq C \leq t\}$ within the subset of women surviving at least to age t.

Since both this risk set and the corresponding event set can be calculated from data restricted to $\mathcal{U}^p$, it is possible to approximate the rates $r(.)$ by $\tilde{r}^*(.)$. Of course, this will be possible only for ages

$$t \geq a^+ \;:=\; \min\{C(u) > 0 \mid u \in \mathcal{U}^p\}$$

which implies that only the conditional survivor function $G(. \mid T \geq a^+)$ can be estimated:

$$G(t \mid T \geq a^+) \;\approx\; \prod_{j=a^+}^{t-1} (1 - \tilde{r}^*(j)) =: G^*(t) \tag{3.13}$$

Notice that in general

$$\mathcal{U}_t^p = \{u \in \mathcal{U}^p \mid T(u) \geq t, 0 \leq C(u) \leq t\} \subsetneq \{u \in \mathcal{U}^p \mid T(u) \geq t\} = \mathcal{U}^p$$

because a women in $\mathcal{U}^p$ might have her first child later than t. In order to create suitable risk sets $\mathcal{U}_t^p$ one has to apply the same conditioning as used for the event sets to meet the assumption that mortality does not depend on whether and when women become mothers.

The following example illustrates the reasoning. Assume that $\mathcal{U}$ refers to a set of 1000 women and consider, in turn, five age classes:

0 In the age class $t = 0$ all 1000 women are at risk of dying, and I assume that 100 women actually die.

1 There remain 900 women. 200 of them will bear a first child in the age class $t = 1$. Also, 100 of the 900 will die. Implied by the assumption that mortality does not depend on becoming a mother, approximately

$$\frac{100}{900}\,200 \approx 22$$

of the mothers will die.

2 There remain 800 women in the age class $t = 2$. I assume that 200 of these women die and 300 women become mothers of a first child. The assumption of equal mortality implies that about 300 / 4 = 75 of these women also die. In addition, there are 178 women who became mothers in age class $t = 1$, and of these about $178/4 \approx 45$ will die.

3 The remaining number of women is $800 - 200 = 600$, and I assume that 200 of these women die in the age class $t = 3$. Again, 200 women bear children for the first time. Consequently, about $200/3 \approx 67$ of these women will die. Of the 358 women who became mothers before $t = 3$, 119 will die.

4 Finally, the remaining 400 women are supposed to die in the oldest age class.

Since in this example the complete data are available, one may calculate the survivor function explicitly:

t	$\lvert\mathcal{U}_t\rvert$	$\lvert\{u \in \mathcal{U}_t \mid T(u) = t\}\rvert$	$r(t)$	$G(t)$
0	1000	100	1/10	1.00
1	900	100	1/9	0.90
2	800	200	1/4	0.80
3	600	200	1/3	0.60
4	400	400	1	0.40

Obviously, the survivor function is simply proportional to the number of women in the risk set. In a next step, I assume that data are only available for $\mathcal{U}^p$, that is, for women who gave birth to at least one child. In the example, there are altogether 200 + 300 + 200 = 700 women. I now perform the same calculations for these women using the risk and event sets as defined above. This can be summarised in the following table:

t	$\lvert\mathcal{U}_t^p\rvert$	$\lvert\{u \in \mathcal{U}_t^p \mid T(u) = t\}\rvert$	$\tilde{r}^*(t)$	$G^*(t)$
0				
1	200	22	0.110	
2	478	120	0.251	0.890
3	558	186	0.333	0.667
4	372	372	1.000	0.445

For $t = 0$, the risk set is empty and a death rate can not be calculated. Consequently, the value of the survivor function for $t = 1$ is not estimable either. For $t > 0$ it is possible, however, to create risk and event sets and calculate corresponding rates $\tilde{r}^*(t)$. And these rates can finally be used to derive the values of the conditional survivor function

$$G(t \mid T \geq a^+) \;=\; \frac{G(t)}{G(a^+)} \;\approx\; G^*(t)$$

where $a^+ = 1$ and $G(a^+) = 0.9$ in the present example. Here, $G^*(.)$ refers to the conditional survivor function estimable from the risk sets $\mathcal{U}_t^p$. Note in particular that the cardinality of risk sets may increase with increasing life times, a feature not encountered with right censored data. In contradistinction to the latter case, the statistical precision of estimates may be rather low for the young ages. This fact may cause severe practical problems of identification and robustness. Some aspects of these difficulties are taken up in the next Section.

How does the "simple device" deal with the current problem? As a first step, again, a probability space is introduced and random variables in parallel with (3.12) are defined:

$$(T, C)\colon \mathcal{U} \times \Omega \longrightarrow \mathcal{T} \times \big(\mathcal{T} \cup \{-1\}\big) \tag{3.14}$$

If $T(u, .)$ was stochastically independent from $C(u, .)$ ($T(u, .) \perp\!\!\!\perp C(u, .)$), then the model restricted to the event $\{0 \leq C(u, .) \leq T(u, .)\}$ is called a *random left truncation model.*[23]

Mandel (2007) provides a vivid discussion of the distinction between truncation and censoring. He states: "Whereas censoring is a model of missing observations ..., truncation is a model of selection bias." (p. 322). Since my setup includes an explicit reference to members of the population and assumes that at least the cardinality of $\mathcal{U}$ is known, the truncation model has the additional consequence that the probability model must be extended to a model that includes at least the population size as a random variable.

For a single (possibly hypothetical) mother u the joint distribution of $T(u, .)$ and $C(u, .)$ conditional on $\{a^+ \leq C(u, .) \leq T(u, .)\}$ (u has at least one child) becomes

$$\begin{aligned}
&\Pr(T(u, .) = t, C(u, .) = s \,|\, a^+ \leq C(u, .) \leq T(u, .)) \\
&= \frac{\Pr(T(u, .) = t) \Pr(C(u, .) = s)}{\Pr(a^+ \leq C(u, .) \leq T(u, .))} \, \mathbb{1}[a^+ \leq s \leq t] \\
&= \frac{\Pr(T(u, .) = t) \Pr(a^+ \leq C(u, .) \leq t)}{\Pr(a^+ \leq C(u, .) \leq T(u, .))}
\end{aligned}$$

[23] More details on the left truncation model are presented in Chapter III.3 of Andersen et al. (1993). They also discuss conditions that are less restrictive than stochastic independence. Keiding/Gill (1990) re-parametrise the left truncation model as a three state Markov model, enabling them to provide another relaxation of independence.

$$\times \frac{\Pr(C(u,.)=s)}{\Pr(a^+ \leq C(u,.) \leq t)} \mathbb{1}[a^+ \leq s \leq t]$$

$$= \frac{\Pr(T(u,.)=t)\Pr(C(u,.) \leq t \,|\, C(u,.) \geq a^+)}{\Pr(C(u,.) \leq T(u,.) \,|\, C(u,.) \geq a^+)}$$

$$\times \Pr(C(u,.)=s \,|\, a^+ \leq C(u,.) \leq t)\, \mathbb{1}[a^+ \leq s \leq t]$$

where $T(u,.) \perp\!\!\!\perp C(u,.)$ is used in the first equation. The first factor in the last line is the distribution of $T(u,.)$ in the coarsened model, the second one the conditional distribution of $C(u,.)$ given $\{T(u,.)=t\} \cap \{a^+ \leq C(u,.) \leq T(u,.)\}$. This suggests a simple estimation strategy: The first factor can be estimated by the empirical distribution of the observed death times, i.e. from the distribution of $T(u,.)$ in the coarsened data model. Since also the distribution function $\Pr(C(u,.) \leq t \,|\, C(u,.) \geq a^+)$, the distribution of the age at first birth of all women who had at least one child, can be estimated, the empirical distribution of the observed death times $T(u,.)$ weighted by the inverse of the latter distribution function should be a reasonable estimator of the marginal distribution of the death time $T(u,.)$.

While the problem of reference to population members u with $C(u,\omega) = -1$ as discussed in the previous Section persists, there is an additional problem with the "simple device" as presented up to now: The "assumption" of stochastic independence, which by itself is "untestable", is simply and obviously wrong, at least when the model is to be realistic. For then the random variables $(T(u,.), C(u,.))$ must satisfy both $\Pr(C(u,.) \leq T(u,.)) = 1$ (either children are born before the death of their mother or there are no children to that woman) and $\Pr(C(u,.) = -1 \,|\, T(u,.) < a^+) = 1$ (there are no children to women dying before the reproductive age).

It may help here to be more explicit about the distinction of "assumptions" within the "simple device" and assumptions about facts that are to be represented by the model. Suppressing again the reference to individuals, it is certainly possible to produce a joint distribution of a random vector (T, C) such that $T \perp\!\!\!\perp C$ and such that the requirements $\Pr(C \leq T) = 1$ and $\Pr(C = \text{-}1 \,|\, T < a^+)$ hold. One possibility is to simply restrict the range of (T, C) such that the minimum of the range of T is larger than the maximum of the range of C. In that case, T and C may be independent while still fulfilling the consistency requirements. That is, this form of "independence assumption" is not self-contradictory. It is, moreover, "untestable" for any given data set since by the very construction of the observation scheme neither $\{T < a^+\}$ nor $\{C = \text{-}1\}$ are ever observed. Nevertheless, this model of independence is plainly wrong. It is obvious that there are deaths before age a^+ in human populations. And there are deaths during the reproductive ages. Thus, range restrictions on (T, C) are counter to fact. But if the ranges $\mathcal{C}, \mathcal{T}$ of T and C overlap and $|\mathcal{C} \cap \mathcal{T}| \geq 2$, then independence $T \perp\!\!\!\perp C$ forces $\Pr(C > T) > 0$. But giving birth to a child after death is equally counter to fact. While in many cases model assumptions counter to fact do not rule out the usefulness of the models concerned, in the present case the very assumptions, even if taken only as an "approximation", would undermine the very goal of estimating relevant aspects of the distribution of life lengths of a previous generation.

Thus the simple minded reference to stochastic independence will not do. One has to replace the condition of stochastic independence by a dynamic, local condition similar to the dynamic form of Δ-independence used above. In the latter case, the construction of the risk sets $\mathcal{U}_t^p$ (and the derived rates $\tilde{r}^*(t)$) take into account the temporal evolution of membership in $\mathcal{U}^p$. Women become members of $\mathcal{U}^p$ after the birth of her first child. This can be mimicked in a probability model by asking for a dynamic and local form of independence:[24]

$$\begin{aligned}\Pr(T(u,.) = t \mid T(u,.) \geq t)&\\ &= \Pr(T(u,.) = t \mid T(u,.) \geq t \geq a^+, a^+ \leq C(u,.) \leq t)\end{aligned}$$

i.e. $\{T(u,.) = t\} \perp\!\!\!\perp \{a^+ \leq C(u,.) \leq t\}$ given $\{T(u,.) \geq t \geq a^+\}$. This condition might be used to justify the construction of consistent estimators of the conditional survivor function within the framework of the "simple device". However, there is no obvious, non-trivial joint

[24] Further details concerning concepts of dynamic or local independence are discussed in Chapter 5.

distribution of $(T(u,.), C(u,.))$ that obeys this form of local independence. The "simple device", while illuminating in the context of simple (and important) examples, becomes rather cumbersome in the present context.

3.7. Selection by Survival

The problem of "left truncation" is not the last problem that must be tackled before a reasonable use of information from the children generation for inference on the distribution of the life length of the previous generation can be proposed. As already indicated in the introductory remarks to this Chapter, the available information from survey data will depend on the survival of the children up to the interview date. While the problem can be discussed in a similar way as was done up to now, in this Section I will propose a simulation model that simultaneously takes into account the problem of selection by survival and most of the previously discussed problems. The aim is to construct a background against which the "simple device" can be judged. The construction of the simulation model will be based on a probabilistic formulation. It is therefore favourable to the "simple device" and allows for the introduction of further probabilistic elements whose merits for arguments and speculations within the "simple device" can then be judged.

At the same time, I will try to build the many features characteristic for the present problem into the generating procedure. Moreover, data from official sources will be used as templates for the distributions employed for the data generation. This will allow for a comparison of the simulation results with the empirical results from survey data reported later in this Chapter.

3.7.1. The Simulation Model

The basic idea is to simulate data for a set of women according to a known survivor function and then to compare this known function with estimates based on information from the women's children who survived until some fixed interview date. In the first version of the model I refer to a set of $N = 10000$ women all born in the year $t_0 := 1900$; this set will be denoted by $\mathcal{U}$. I assume that these women survive according to the 1891–1900 period life table for Germany (see Table 3.3); the corresponding age-specific death rates will be denoted by δ^f_t. Further, the life tables for their children will be considered. The following table describes the official period life tables available for Germany:

Period	Publication
1871 – 1880	Statistik des Deutschen Reichs, Vol. 246 (pp. 14*-17*).
1881 – 1890	Statistik des Deutschen Reichs, Vol. 246 (pp. 14*-17*).
1891 – 1900	Statistik des Deutschen Reichs, Vol. 246 (pp. 14*-17*).
1901 – 1910	Statistik des Deutschen Reichs, Vol. 246 (pp. 14*-17*).
1910 – 1911	Statistik des Deutschen Reichs, Vol. 275. Statistisches Jahrbuch für das Deutsche Reich 1919 (pp. 50-51).
1924 – 1926	Statistik des Deutschen Reichs, Vol. 360 and 401. Statistisches Jahrbuch für das Deutsche Reich 1928 (pp. 38-39).
1932 – 1934	Statistik des Deutschen Reichs, Vol. 495 (pp. 86-87). Statistisches Jahrbuch für das Deutsche Reich 1936 (pp. 45-46).
1949 – 1951	Statistik der Bundesrepublik Deutschland, Vol. 75 and 173. Statistisches Jahrbuch für die Bundesrepublik Deutschland 1954 (pp. 62-63).
1960 – 1962	Statistisches Jahrbuch für die Bundesrepublik Deutschland 1965 (pp. 67-68).
1970 – 1972	Fachserie 1, Reihe 2, Sonderheft 1. Allgemeine Sterbetafel für die Bundesrepublik Deutschland 1970/72.
1986 – 1988	Fachserie 1, Reihe 1, Sonderheft 2. Allgemeine Sterbetafel für die Bundesrepublik Deutschland 1986/88.

All tables are period life tables. Until 1932 – 34, they refer to the territory of the former *Deutsches Reich*; all other tables refer to the territory of the former FRG. Methods of table construction have slightly changed throughout the years.

The following table 3.3 gives the surviving number of women at age *t* from these life tables.

Table 3.3.: Female survivor functions. German period life tables.

t	1871/ 1881	1881/ 1890	1891/ 1900	1901/ 1910	1910/ 1911	1924/ 1926	1932/ 1934	1949/ 1951	1960/ 1962	1970/ 1972	1986/ 1988
0	100000	100000	100000	100000	100000	100000	100000	100000	100000	100000	100000
1	78260	79311	80138	82952	84695	90608	93161	95091	97222	98016	99298
2	73280	74404	76137	79761	82070	89255	92394	94749	97027	97888	99241
3	70892	72073	74482	78594	81126	88743	92026	94545	96922	97810	99201
4	69295	70514	73406	77867	80523	88422	91761	94390	96845	97745	99174
5	68126	69377	72623	77334	80077	88169	91535	94270	96782	97690	99153
6	67249	68537	72038	76924	79730	87975	91338	94177	96728	97641	99136
7	66572	67881	71577	76587	79445	87817	91160	94100	96682	97597	99119
8	66035	67358	71206	76301	79206	87683	91003	94041	96643	97558	99103
9	65599	66942	70903	76058	79001	87563	90870	93986	96609	97523	99088
10	65237	66601	70646	75845	78816	87452	90753	93937	96579	97492	99073
11	64926	66309	70420	75651	78642	87347	90650	93893	96552	97465	99058
12	64649	66049	70210	75467	78476	87243	90557	93850	96525	97439	99044
13	64390	65801	70003	75285	78311	87134	90467	93805	96498	97413	99029
14	64136	65555	69789	75094	78131	87013	90373	93756	96468	97384	99013
15	63878	65306	69562	74887	77930	86877	90270	93701	96434	97349	98995
16	63609	65045	69319	74661	77710	86719	90152	93637	96395	97305	98974
17	63322	64764	69060	74411	77470	86534	90016	93564	96351	97251	98947
18	63013	64468	68787	74143	77216	86319	89858	93484	96301	97189	98916
19	62681	64160	68500	73861	76945	86075	89680	93394	96246	97124	98881
20	62324	63838	68201	73564	76659	85808	89490	93295	96188	97059	98843
21	61941	63500	67888	73254	76362	85523	89287	93188	96128	96996	98806
22	61534	63142	67559	72929	76052	85226	89072	93073	96068	96934	98768
23	61102	62762	67212	72586	75730	84920	88849	92955	96008	96874	98731
24	60648	62360	66848	72225	75397	84602	88622	92834	95948	96815	98694
25	60174	61937	66467	71849	75043	84275	88390	92711	95884	96755	98657
26	59680	61497	66072	71463	74668	83943	88151	92586	95814	96694	98619
27	59170	61042	65666	71070	74283	83610	87904	92457	95739	96632	98579
28	58647	60570	65249	70669	73896	83274	87653	92324	95660	96567	98538
29	58111	60082	64822	70261	73513	82937	87397	92185	95575	96499	98493
30	57566	59584	64385	69848	73115	82597	87139	92039	95485	96429	98446
31	57010	59076	63937	69432	72703	82254	86876	91887	95390	96355	98395
32	56445	58554	63479	69008	72291	81909	86607	91729	95290	96276	98340
33	55869	58018	63010	68575	71876	81559	86329	91565	95184	96190	98280
34	55282	57473	62533	68132	71457	81205	86044	91396	95071	96098	98216
35	54685	56921	62047	67679	71020	80847	85754	91221	94949	95997	98146
36	54078	56360	61549	67215	70554	80482	85455	91039	94818	95886	98071
37	53462	55789	61041	66744	70080	80105	85145	90850	94676	95764	97988
38	52837	55215	60524	66266	69610	79720	84819	90651	94524	95632	97896
39	52207	54638	59998	65779	69139	79324	84481	90443	94360	95488	97796
40	51576	54054	59467	65283	68659	78917	84135	90225	94184	95331	97685

Table 3.3.: Female survivor functions. German period life tables.

t	1871/ 1881	1881/ 1890	1891/ 1900	1901/ 1910	1910/ 1911	1924/ 1926	1932/ 1934	1949/ 1951	1960/ 1962	1970/ 1972	1986/ 1988
41	50946	53467	58931	64779	68172	78498	83779	89995	93995	95161	97564
42	50320	52880	58391	64269	67689	78068	83410	89749	93792	94975	97431
43	49701	52297	57848	63754	67194	77627	83027	89486	93573	94773	97286
44	49090	51720	57302	63238	66692	77175	82630	89204	93337	94551	97127
45	48481	51146	56751	62717	66187	76704	82211	88901	93081	94308	96954
46	47870	50569	56195	62181	65661	76210	81763	88574	92803	94042	96766
47	47248	49983	55628	61628	65105	75688	81282	88221	92500	93750	96562
48	46605	49385	55040	61053	64510	75136	80767	87841	92173	93427	96341
49	45939	48765	54423	60449	63883	74557	80213	87432	91821	93072	96102
50	45245	48110	53768	59812	63231	73943	79620	86991	91442	92683	95842
51	44521	47418	53078	59138	62547	73289	78990	86516	91035	92260	95559
52	43767	46692	52354	58418	61827	72592	78322	86003	90597	91806	95252
53	42981	45934	51594	57648	61048	71854	77613	85451	90125	91323	94918
54	42162	45136	50791	56837	60219	71071	76855	84860	89615	90813	94553
55	41308	44293	49938	55984	59350	70236	76038	84225	89063	90272	94156
56	40414	43396	49032	55077	58441	69342	75162	83540	88464	89696	93723
57	39472	42448	48072	54106	57468	68383	74225	82796	87814	89078	93252
58	38476	41462	47054	53067	56398	67357	73221	81989	87105	88411	92738
59	37418	40415	45971	51959	55245	66257	72142	81115	86331	87689	92179
60	36293	39287	44814	50780	54016	65076	70984	80166	85484	86903	91569
61	35101	38087	43582	49524	52713	63809	69745	79131	84556	86044	90903
62	33843	36823	42272	48176	51320	62448	68409	77994	83538	85101	90178
63	32521	35497	40880	46725	49816	60973	66960	76744	82420	84062	89387
64	31140	34102	39398	45178	48199	59377	65396	75374	81191	82915	88526
65	29703	32628	37828	43540	46484	57671	63712	73875	79839	81647	87587
66	28217	31088	36179	41816	44693	55852	61895	72232	78352	80250	86565
67	26686	29506	34460	40007	42782	53901	59933	70428	76720	78713	85451
68	25118	27897	32675	38111	40773	51813	57822	68455	74932	77027	84236
69	23521	26252	30826	36129	38663	49597	55568	66312	72976	75179	82909
70	21901	24546	28917	34078	36448	47255	53184	63994	70840	73157	81459
71	20265	22786	26956	31963	34191	44799	50652	61491	68513	70948	79869
72	18617	21000	24957	29777	31830	42248	47951	58794	65981	68539	78124
73	16960	19204	22938	27535	29379	39609	45118	55905	63235	65920	76206
74	15307	17416	20914	25273	26933	36869	42182	52837	60267	63084	74096
75	13677	15645	18900	23006	24517	34024	39132	49605	57076	60033	71775
76	12090	13892	16919	20745	22106	31126	35989	46226	53674	56774	69230
77	10569	12219	15000	18526	19673	28217	32820	42721	50082	53323	66447
78	9131	10661	13163	16372	17336	25335	29670	39118	46331	49702	63419
79	7795	9192	11417	14299	15112	22487	26559	35457	42458	45934	60148
80	6570	7815	9773	12348	12981	19711	23500	31787	38507	42046	56640
81	5464	6550	8252	10539	11016	17075	20527	28163	34529	38076	52912
82	4479	5408	6869	8864	9184	14624	17691	24642	30579	34071	48992
83	3614	4394	5626	7329	7499	12353	15026	21282	26717	30091	44916
84	2867	3511	4524	5955	6030	10262	12561	18132	23004	26204	40734
85	2232	2756	3568	4752	4794	8372	10323	15225	19500	22478	36501
86	1705	2124	2764	3719	3746	6712	8324	12582	16258	18974	32282
87	1276	1605	2104	2850	2856	5290	6567	10213	13319	15744	28146
88	935	1189	1571	2138	2140	4101	5075	8132	10705	12826	24160

Table 3.3.: Female survivor functions. German period life tables.

t	1871/ 1881	1881/ 1890	1891/ 1900	1901/ 1910	1910/ 1911	1924/ 1926	1932/ 1934	1949/ 1951	1960/ 1962	1970/ 1972	1986/ 1988
89	671	862	1149	1571	1574	3128	3857	6335	8147	10245	20393
90	471	612	821	1131	1126	2356	2868	4815	6480	8016	16903
91	323	424	573	797	786	1736	2083	3567	4872	6139	13738
92	217	288	390	549	534	1256	1476	2571	3580	4597	10935
93	142	191	260	370	354	891	1019	1814	2571	3362	8511
94	90	123	169	244	228	620	683	1253	1805	2409	6468
95	56	78	107	157	142	423	445	846	1240	1671	4792
96	34	48	66	99	87	283	281	559	834	1134	3457
97	20	29	40	61	51	185	172	361	550	750	2425
98	11	17	24	38	29	119	101	227	356	483	1651
99	6	10	14	22	16	74	58	140	227	303	1090
100	3	6	8	13	9	45	31	84	142	185	697

Additionally, the birth of children must be considered. While women as well as men may become parents in a legal sense in many different ways, I will only consider women who gave birth to children. Age- and parity-specific birth rates are given by

$$\beta_{t,k} := \frac{\text{Number of women giving birth to a further child at age } t}{\text{Number of women aged } t \text{ and having } k \text{ children}}$$

In order to arrive at a simulation model that roughly corresponds to the historical situation these rates are calculated from a subsample of the census that took place in Germany in the year 1970.[25]

The census of 1970 was conducted on May 27 of that year in the territory of the former FRG.[26] As part of this census a subsample of 10% of the population was asked to provide additional information, in particular, all women with a German citizenship who participated in the 10% subsample were asked for dates of marriage and birth dates of all their marital children, regardless of their current marital status.[27] Several

[25] The birth rates normally reported by the statistical office refer only to marital births, starting parity counts afresh with each marriage. The non- or under-reporting of births per woman will at least continue to the next planned census in 2011.

[26] For a detailed description, including a presentation of the questionnaire see Schubnell and Herberger (1970).

[27] Some results from these additional questions were published, albeit in highly aggregated form, by the *Statistisches Bundesamt* in Fachserie A.

years ago, official statistics in Germany agreed to make available an anonymised 10% subsample of the 1970 census.[28] This 1% subsample of all women who lived in May 1970 in the territory of the former FRG and had a German citizenship is used in the sequel. The number of cases is 314993. If multiplied by 100, this is roughly the number of women with a German citizenship living in the former FRG in May 1970.

For the calculation of age- and parity-specific birth rates I have used all of the women born between 1870 and 1925, altogether 131135 cases.[29] While these rates may not be very accurate, are spanning very different historical contexts, and are themselves subject to selection- and migration problems (including possible differential mortality due to childbearing etc.), the main issue here is to provide a historically reasonable background to judge statistical methods, a task that may be accomplished by these numbers. An accurate account of the demography of the period is not intended here.

Table 3.4.: Parity-specific birth rates per 10000 women

t	1	2	3	4	5	6	7	8	9	10	11	12
16	10	286	0	0	0	0	0	0	0	0	0	0
17	39	185	0	0	0	0	0	0	0	0	0	0
18	102	432	99	0	0	0	0	0	0	0	0	0
19	196	628	318	769	0	0	0	0	0	0	0	0
20	332	885	644	482	2000	0	0	0	0	0	0	0
21	455	999	776	651	0	0	0	0	0	0	0	0
22	598	1123	902	905	897	1250	0	0	0	0	0	0
23	702	1245	1040	1128	737	870	0	0	0	0	0	0
24	800	1299	1073	988	1045	923	2222	0	0	0	0	0
25	860	1317	1109	1090	897	1260	1304	4000	0	0	0	0
26	911	1278	1088	1121	1167	1133	816	1429	3333	0	0	0
27	905	1304	1009	1138	1238	1148	1237	556	0	10000	0	0
28	883	1255	1028	1150	1307	1395	1230	500	2000	0	0	0
29	844	1224	975	1103	1179	1589	954	1884	588	0	0	0
30	753	1176	922	1074	1263	1398	1835	1172	1818	1250	0	0
31	679	1078	886	1004	1191	1384	1442	1682	2031	1000	0	0

[28] More information on these data sets are available from the *Zentrum für Umfragen, Methoden und Analysen* (ZUMA, Mannheim), Abteilung für Mikrodaten; see: http://www.gesis.org/Dauerbeobachtung/GML.

[29] I thank Bernhard Schimpl-Neimanns (ZUMA) who prepared the basic counts used here. I still had to delete one woman with a negative age at birth.

Table 3.4.: Parity-specific birth rates per 10000 women

t	1	2	3	4	5	6	7	8	9	10	11	12
32	597	1016	792	890	1177	1320	1675	2143	1333	2059	2727	0
33	510	937	740	914	1071	1304	1371	1628	2262	1231	0	5000
34	433	823	675	834	1074	1326	1480	1470	2087	1800	1515	1667
35	355	711	610	726	997	1198	1242	1587	1870	1985	1852	667
36	305	588	537	692	863	1133	1500	1476	1280	1905	2222	2000
37	259	499	466	626	772	993	1136	1529	1408	2247	1875	2619
38	197	373	422	523	713	914	998	1520	1574	1290	1981	2115
39	164	315	316	444	613	792	911	1352	1565	1486	1587	1803
40	122	243	245	355	510	642	868	1383	1126	1447	1600	1486
41	91	154	187	275	439	544	782	1017	1057	851	1494	930
42	59	102	133	179	275	451	539	765	930	1212	1371	1667
43	39	64	83	119	200	309	435	581	747	732	787	808
44	25	43	51	81	130	177	231	459	389	543	761	1143
45	14	20	28	48	60	141	159	214	255	548	316	707
46	8	12	11	23	27	42	62	122	211	28	432	400
47	7	7	6	12	20	23	31	47	31	56	0	306
48	5	3	4	6	14	12	13	9	15	84	110	0
49	2	3	3	2	3	5	9	9	15	0	0	0
50	1	1	1	1	2	5	18	0	0	0	0	0

From a starting set of identifying numbers $\mathcal{U}$ of women born in 1900, their children are represented by a second set $\mathcal{U}^c := \cup_{u \in \mathcal{U}} c(u)$ containing identification numbers of all children born to the women in $\mathcal{U}$. The number of members of $\mathcal{U}^c$ is not known in advance but depends on both the death rates δ_t^f and the birth rates $\beta_{t,k}$. Given these rates, I created two lists. One list contains for each woman in $\mathcal{U}$ her identification number, her death year, and her number of children. Another list contains for each child in $\mathcal{U}^c$ an identification number, the birth year, and the identification number of the mother. In addition, in order to simulate a retrospective survey, I assume that the children survive according to the 1960–1962 period life table for women (see Table 3.3); the corresponding age-specific death rates will be denoted by δ_t^c. Using these rates, a simulated death year is added for each child in the second list.

Box 3.7-1 summarises the algorithm used to generate the simulated data. In the description, the $\epsilon_.$ refer to draws from random numbers uniformly distributed in the unit interval. Using this algorithm I get a first list with

Box 3.7-1 Skeleton of the simulation model.

```
For each $u \in \mathcal{U}$ do:
    $n(u) := 0$;                    # counter for $u$'s children
    For ($t = 0, \ldots, 100$) {
        Generate a new, independent random number $\epsilon_{t,b}(u)$;
        If ($\epsilon_{t,b}(u) \leq \beta_{t,n(u)}$)
          increase $n(u)$ by 1; create a new entry in $\mathcal{U}^c$,
          and record the mother's identification number and age;
        Get, independently, another random number $\epsilon_{t,d}(u)$;
        If ($\epsilon_{t,d}(u) \leq \delta_t^f$)
          goto L1;
    }
L1:
    Record that $u$ died at age $t$ and has given birth to $c(u)$
    children, also record for all children the mother's age at death;

For each $u \in \mathcal{U}^c$ do:            # Simulate childrens survival
    For ($t = 0, \ldots, 100$) {
        Get another, independent random number $\epsilon_{t,c}(u)$;
        If ($\epsilon_{t,c}(u) \leq \delta_t^c$)
          goto L2;
    }
L2:
    Record that $u$ died at age $t$;
```

$N = 10000$ entries that records the identification numbers of the women, $\mathcal{U}$, their age at death, and their number of children. In one run of the algorithm, presented below, 4776 of these 10000 women have at least one child.[30] The second list contains entries for 11407 children. Figure 3.6 shows a frequency distribution of the years in which the women in $\mathcal{U}$ died on a historical time axis. Also shown are frequency distributions of the birth and death years of the children. Note that the algorithm is based on the assumption that women's survival is locally independent of their giving birth to children. The algorithm provides an operational

[30] Note that according to the 1891–1900 period life table, only 68% of the women survived age 20, and only 60% survived age 40.

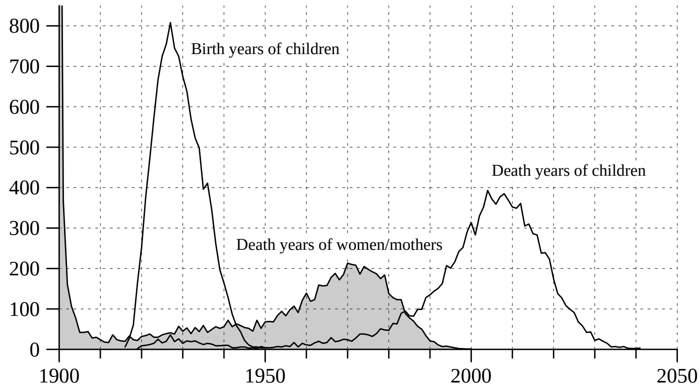


Figure 3.6.: Frequency distributions of birth and death years in the simulated data set.

explication of this notion in discrete time: given survival to age t, the draws of random numbers $\epsilon_{t,b}$ and $\epsilon_{t,d}$ used to simulate births and deaths at age t are stochastically independent. [31]

3.7.2. Considering Left Truncation

Before using the model to discuss the question of how to recover the survivor function of the members of $\mathcal{U}$ based on information resulting from a retrospective survey of their children, I illustrate the importance

[31] Problems arising from a violation of this assumption can not be checked within this model. However, it can be adapted rather easily to allow for some form of (local) dependence. Also, dependence of life lengths of mothers and children (both local and global) might be incorporated.
It is of some theoretical interest to investigate whether a simulation algorithm may be conceived that does not involve a fixed starting population $\mathcal{U}$ of potential mothers. After all, the surveyor will never have access to that population nor even direct information about its size. Such an algorithm ought to be able to reproduce all reasonable distributions of the life lengths of the previous generation without any extraneous assumptions on the dependence of life length across generations.

of correctly taking into account left truncated observations and discuss some complications arising from real data limitations.

The goal is to recover (some part of) the distribution of the statistical variable

$$T : \mathcal{U} \longrightarrow \mathcal{T} := \{0, 1, \dots, \tau\}$$

that records the life length of the members of $\mathcal{U}$ from observations on a selected subset of $\mathcal{U}$. The subset is not determined by some random device under the control of the survey administrator or some process that may be judged similar to such a device. Instead, the observations refer to the subset of members of $\mathcal{U}$ having given birth to at least one child. This subset will be denoted by $\mathcal{U}^p$. In parallel to the definition of T one may define a statistical variable

$$T^* : \mathcal{U}^p \longrightarrow \mathcal{T}$$

that records the life length of the members of $\mathcal{U}^p$. Then the statistical variable T^* is the restriction of the function $T : \mathcal{U} \rightarrow \mathcal{T}$ to the subset $\mathcal{U}^p$.

In the simulation model of the previous Section, the distribution of T is defined by the death rates δ_t^f. The survivor function of T is therefore given by

$$G(t) = \prod_{j=0}^{t-1} (1 - \delta_j^f)$$

One possibility to recover (some part of) this survivor function is to somehow estimate the death rates δ_t^f. However, these death rates may be systematically different from the naively computed death rates derived from the statistical variable T^*

$$r^*(t) := \frac{|\{u \in \mathcal{U}^p \mid T^*(u) = t\}|}{|\{u \in \mathcal{U}^p \mid T^*(u) \geq t\}|} = \mathrm{P}(T^* = t \mid T^* \geq t)$$

The rate function $r^*(.)$ refers to deaths among those women who ever became mothers. But in order to find estimates of δ_t^f, one needs to take into account that women only become members of $\mathcal{U}^p$ when they have given birth to a first child and not on the fact that they ever gave birth.

Recalling the framework of Section 3.6, consider the two-dimensional variable

$$(T, C) \colon \mathcal{U} \longrightarrow \mathcal{T} \times \big(\mathcal{T} \cup \{-1\}\big)$$

As before, $T(u)$ denotes u's life length and $C(u)$ refers to the age at which u became for the first time mother of a child. If u stays childless throughout her lifetime, $C(u)$ is set to -1. In a temporal perspective, therefore, the value of $C(u)$ is only determined after u's death.

Remember further the definitions of the temporarily evolving populations

$$\mathcal{U}_t \coloneqq \{u \in \mathcal{U} \mid T(u) \geq t\} \text{ and } \mathcal{U}_t^p \coloneqq \{u \in \mathcal{U}^p \mid T(u) \geq t, 0 \leq C(u) \leq t\}$$

One may then define the rate function

$$\tilde{r}^*(t) \coloneqq \frac{|\{u \in \mathcal{U}_t^p \mid T(u) = t\}|}{|\mathcal{U}_t^p|} = \mathrm{P}(T = t \mid T \geq t, 0 \leq C \leq t)$$

which refers to the death rates in the population of mothers conditional on a previous birth.

Generally, $\tilde{r}^*(t) \geq r^*(t)$, and thus using $r^*(.)$ underestimates the rate and the survivor function is overestimated. This can be seen by comparing the denominators in the definitions of $r^*(.)$ and $\tilde{r}^*(.)$ (the numerators are equal):

$$|\{\mathcal{U}^p \mid T(u) \geq t, 0 \leq C(u) \leq t\}| \leq |\{\mathcal{U}^p \mid T(u) \geq t\}|$$

Now $\tilde{r}^*(t)$ is the death rate of women who actually are members of $\mathcal{U}^p$ *at the age of* t and, as I have construed the model, it is a reasonable estimate of δ_t^f, at least in the extended probability model: The random variables corresponding to (T, C) are

$$(T, C) \colon \mathcal{U} \times \Omega \longrightarrow \mathcal{T} \times \big(\mathcal{T} \cup \{-1\}\big)$$

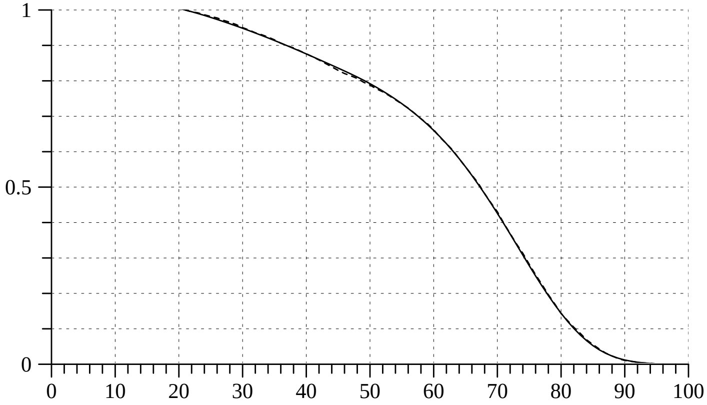


Figure 3.7.: Survivor functions, conditional on $t \geq 21$, from the 1891–1900 period life table (solid line) and from women with at least one child in the simulated data set (dotted line).

The simulation proceeds by stepping forward through time. At each age t, the process ends if $\epsilon_{t,d}(u, \omega) \leq \delta_t^f$ and $T(u, \omega)$ is set to t, otherwise it proceeds to the next age $t + 1$. Thus the simulation provides realisations of life lengths $T(u, .)$ with $\Pr(T(u, .) = t \mid T(u, .) \geq t) = \delta_t^f$ and $\Pr(T(u, .) \geq t) = G(t)$ for all u. With $\mathcal{T}$ finite and $\mathcal{U}$ large, 'most' empirical distribution functions derived from the simulations can be expected to be close to $G(.)$.

The simulation does not directly produce a random variable $C(u, .)$. In order to use the results also for the discussion of more complicated situations, it proceeds to simulate data for all births of u. However, the random variable $C(u, .)$ as defined above is a simple function of the random variables $n(u, .)$ and the birth dates recorded in the list of children. And since the simulation for any $u \in \mathcal{U}$ stops with a death event, $C(u, .) \leq T(u, .)$ is ensured. Furthermore, also by construction

$$\Pr(T(u, .) = t \mid T(u, .) \geq t) = \Pr(T(u, .) = t \mid T(u, .) \geq t, 0 \leq C(u, .) \leq t)$$

But the former quantity is δ_t^f and the latter quantity is the one estimated by $\tilde{r}^*(.)$ so that $\tilde{r}^*(.)$ should be close to δ_t^f in reasonably large samples.

In principle, it would be possible to obtain estimates $\tilde{r}^*$ from the earliest observed age at first birth onward. Since this rate is zero up to the age of the first observed death in $\mathcal{U}^p$, one might as well start at this age, say a^+. On the other hand, due to the small number of cases in a sample of observations, $\tilde{r}^*(a^+)$ might not be a good estimate of $\delta^f_t(a^+)$ and one should condition on some later age. In fact, it might even happen that $\tilde{r}^*(a^+) = 1$ so that one cannot find a reasonable estimate of a survivor function beginning at a^+.

Having chosen a reasonable a^+, the conditional survivor function derived from the estimates $\tilde{r}^*(.)$ becomes

$$\tilde{G}^*_{a^+}(t) := \prod_{j=a^+}^{t-1} (1 - \tilde{r}^*(j))$$

It might be taken as an estimate of the conditional survivor function $G(. \,|\, T \geq a^+)$. In the simulated data set, the earliest death occurs at age 19. However, this occurs only once, and at $t = 20$ there is no death at all. I therefore define $a^+ := 21$. The estimate $\tilde{G}^*_{a^+}(.)$ is shown in Figure 3.7 as a dotted line together with $G(. \,|\, T \geq a^+)$ calculated from the 1891-1900 period life table for women. Obviously, both curves agree quite well. On the other hand, if I had not taken into account the fact that women become members of $\mathcal{U}^p$ only after having given birth to a first child, but estimated the survivor function of T^*, the result would be systematically biased as a consequence of the inequality $r^*(t) \leq \tilde{r}^*(t)$. This is illustrated by Figure 3.8 where the solid line shows $\tilde{G}^*_{a^+}$ and the dotted line shows $\mathrm{P}(T^* \geq . \,|\, T^* \geq a^+)$.

The fact that women become members of $\mathcal{U}^p$ only after the birth of a child is formally equivalent to treating the observations as left truncated at the age at first birth. Of course, nothing is wrong with estimating the survivor function of T^* instead of using $\tilde{G}^*_{a^+}(.)$. The argument has only shown that one should use the latter if the interest is in recovering part of the distribution of T. One might also notice that, while T^* refers to a well-defined statistical variable, there is no statistical variable of which $\tilde{G}^*_{a^+}(.)$ is a conditional survivor function. I.e., the definition of $\tilde{G}^*_{a^+}(.)$ (based on $\tilde{r}^*(.)$) refers to a risk set that changes not only in

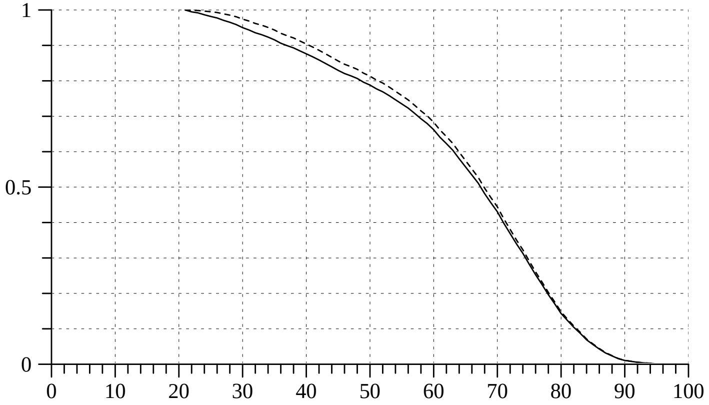


Figure 3.8.: Conditional survivor functions $\tilde{G}^*_{a^+}(.)$ (solid line) and $\mathrm{P}(T^* \geq .\,|\,T^* \geq a^+)$ calculated from the simulated data set with $a^+ = 21$.

accordance with the mortality of (a fixed subset of) women. It also depends dynamically on the fertility of the women. In fact, $\tilde{r}^*(.)$ is a mixture of rate functions defined for subsets of $\mathcal{U}^p$. To see this, partition $\mathcal{U}^p$ into subsets

$$\mathcal{U}^p_{[a]} := \{u \in \mathcal{U}^p \,|\, C(u) = a\} = C^{-1}(\{a\})$$

consisting of those members of $\mathcal{U}^p$ who had a first birth at the age a. The mortality experience of this subset can be summarised by the rate function

$$\tilde{r}^*_a(t) := \frac{|\{\mathcal{U}^p_{[a]} \,|\, T(u) = t\}|}{|\{\mathcal{U}^p_{[a]} \,|\, T(u) \geq t\}|} = \mathrm{P}(T = t \,|\, T \geq t, C = a)\ \mathbb{1}[t \geq a]$$

The rate function $\tilde{r}^*(.)$ arises as a mixture of the rate functions $\tilde{r}^*_a(.)$:

$$\tilde{r}^*(t) = \sum_{a \leq t} \tilde{r}^*_a(t)\, w_a(t)$$

where the weights, defined as

$$w_a(t) := \mathrm{P}(C = a \,|\, C \leq t, T \geq t)\ \mathbb{1}[t \geq a]$$

reflect the composition of the risk set at t. That there is in general no fixed subset of women in $\mathcal{U}^p$ whose mortality experience is 'representative' of the mortality experience of $\mathcal{U}$ reflects the difficulties created by the dynamic selection of $\mathcal{U}^p$.

3.7.3. Using Information from Children

Turning to the question of how to estimate conditional survivor functions for the members of $\mathcal{U}$ when only information from their children (members of $\mathcal{U}^c$) is available, one needs to take into account the relationship between $\mathcal{U}^c$ and $\mathcal{U}^p$. Recall from Section 3.1 the definitions of a few functions that describe generational dependencies:

$$m : \mathcal{U}^c \longrightarrow \mathcal{U}^p \cap \mathcal{U}^f$$

where for each child $u \in \mathcal{U}^c$, $m(u)$ refers to the mother of u in $\mathcal{U}^p$. Conversely, for each woman $u \in \mathcal{U}^p$, $m^{-1}(\{u\})$ is the set of her children in $\mathcal{U}^c$.

Now let s denote a simple random sample from $\mathcal{U}^c$. This induces a random sample from $\mathcal{U}^p$, namely

$$s^* := \{u \in \mathcal{U}^p \mid \text{there is an } u' \in s \text{ with } m(u') = u\}$$

But s^* is not a simple random sample from $\mathcal{U}^p$ because women with more children are more likely to be included in s^*. This should be taken into account when estimating $\tilde{r}^*(t)$ from information provided by the children in the sample s.

The further problem, considered in Section 3.7.2, of the temporal nature of the membership of women in $\mathcal{U}^p$, can be traced from survey responses of children if the variable C referring to the age at the first birth of the mothers (alternatively, the birth dates of all siblings) is recorded. Accordingly, if u is included in s, one should not condition on her mother's age when giving birth to u, but on the age of her mother's first child-bearing. To illustrate the difference, I use the data from the simulation model and compare two fictitious samples: s_1 contains all first-born children from $\mathcal{U}^c$, and s_2 contains all last-born children from

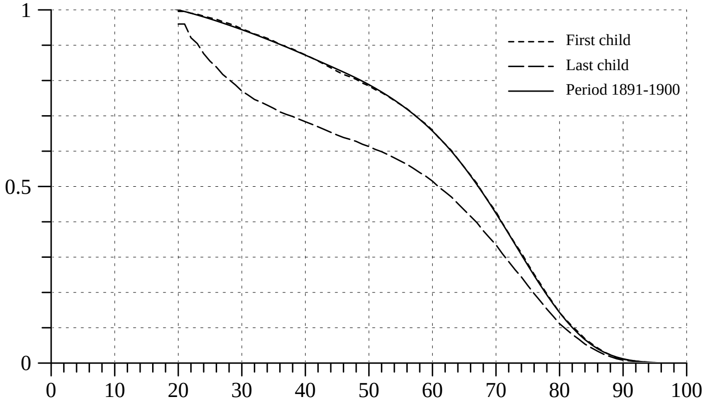


Figure 3.9.: Comparison of conditional survivor functions calculated from two different samples from the simulated data set.

$\mathcal{U}^c$. Of course, both samples provide the same information about the life length of women in $\mathcal{U}^p$. But there are now different ways to select truncation times. If I condition on the age of the mothers when giving birth to the children in the samples, I get the results shown in Figure 3.9. Obviously, conditioning on the mother's age when giving birth to her last child would result in an extremely biased estimate.

In order to avoid this mistake, ideally, the following statistical variables should be available from a sample of children:

$$(T, C, N) \colon s^* \subseteq \mathcal{U} \longrightarrow \mathcal{T} \times \mathcal{T} \times \{1, 2, 3, \ldots\}$$

where $T(m(u))$ records the life length of u's mother, $C(u)$ denotes mother's age at *first* child-bearing, and $N(u)$ counts the mother's number of children. Since N will be used to provide weights for the observations in the sample s^*, this should be the number of children surviving up to the time when the sample is drawn. Assuming that this information is available from a simple random sample s and in conformance with the discussion in Section 3.3 (now restricted to mothers), the rates $\tilde{r}^*$ can be estimated by:

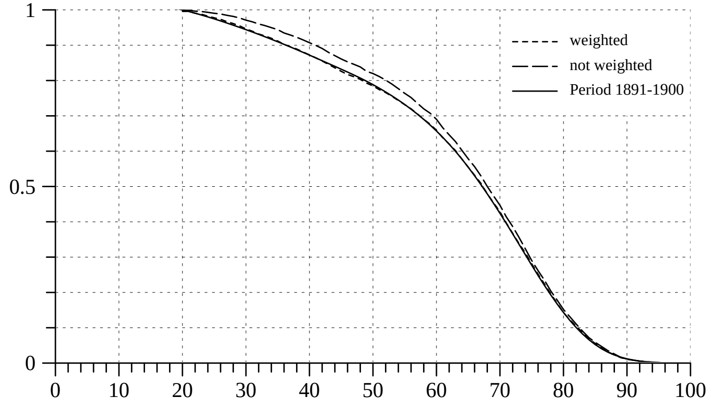


Figure 3.10.: Comparison of conditional survivor functions estimated with, and without, weights from the simulated data set.

$$\tilde{r}^*(t) \approx \tilde{r}^*_w(t) := \frac{\sum_{u \in s} \frac{1}{N(u)} \, \mathbb{1}[T = t, C \leq t](u)}{\sum_{u \in s} \frac{1}{N(u)} \, \mathbb{1}[T \geq t, C \leq t](u)}$$

To illustrate, I use again data from the simulation model. Figure 3.10 compares conditional survivor functions calculated from estimated rates $\tilde{r}^*_w(t)$ and from analogously defined rates where the weights are dropped.[32] The figure clearly indicates that one should use the weights $1/N$ if this information is available.

However, general surveys often do not provide this information. It is therefore important to explore another way to arrive at reasonable estimates. In order to explain this possibility consider the risk set

$$\mathcal{U}^p_t = \{u \in \mathcal{U}^p \mid T(u) \geq t, C(u) \leq t\}$$

[32] In the calculation I have used all observations from $\mathcal{U}^c$, but basically the same differences would result from a simple random sample from $\mathcal{U}^c$.

at t. The death rates to be estimated can then be written as

$$\tilde{r}^*(t) = \frac{|\{u \in \mathcal{U}_t^p \mid T(u) = t\}|}{|\mathcal{U}_t^p|} = \mathrm{P}(T = t \mid T \geq t, C \leq t) \approx \delta_t^f$$

By assumption, these rates do not depend on the number of children born of members of $\mathcal{U}_t^p$ until t, and also do not depend on the children's birth dates. To make this explicit, partition the risk sets into subsets according to the number of children born until t. Let $K_t^*(u)$ denote the number of children born of u until t. Each risk set $\mathcal{U}_t^p$ may then be written as a union of subsets

$$\mathcal{U}_{t,k}^p \coloneqq \{u \in \mathcal{U}_t^p \mid K_t^*(u) = k\}$$

taken over all possible values of k. Furthermore, I can define death rates for these subsets,

$$\tilde{r}_k^*(t) \coloneqq \frac{|\{u \in \mathcal{U}_{t,k}^p \mid T^*(u) = t\}|}{|\mathcal{U}_{t,k}^p|}$$

However, by assumption these rates are all (approximately) identical to the death rate $\tilde{r}^*(t)$. Consequently, I do not need weights when I only use information from children born until t. Instead, I can directly refer to the sets of children born of women in $\mathcal{U}_{t,k}^p$ which can be defined by

$$\mathcal{U}_{t,k}^c \coloneqq m^{-1}(\mathcal{U}_{t,k}^p)$$

The death rates $\tilde{r}_k^*(t)$ may then be written as

$$\tilde{r}_k^*(t) \approx \frac{|\{u \in \mathcal{U}_{t,k}^c \mid T_c^*(u) = t\}|}{|\mathcal{U}_{t,k}^c|}$$

and, since these rates are approximately identical across the subsets, I might finally write

$$\tilde{r}^*(t) \approx \tilde{r}_c^*(t) \coloneqq \frac{|\{u \in \mathcal{U}^c \mid T_c^*(u) = t, S_c^*(u) \leq t\}|}{|\{u \in \mathcal{U}^c \mid T_c^*(u) \geq t, S_c^*(u) \leq t\}|} \tag{3.15}$$

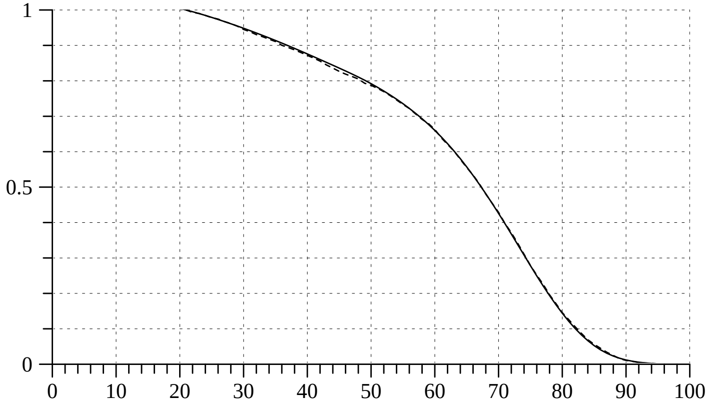


Figure 3.11.: Conditional survivor function estimated from the rates $\tilde{r}_c^*(t)$, compared with a conditional survivor function from the 1891-1900 period life table.

where now $S_c^*(u)$ is the age of u's mother at the birth of u. Notice that this approach does not require any weights and also requires no information about the mothers age at her first child-bearing.

To illustrate the argument I use again data from the simulation model. Taking into account all children in $\mathcal{U}^c$ but, for the calculation of the rates $\tilde{r}_c^*(t)$ only use information from children born not later than t. Of course, this simply means to use all information from $\mathcal{U}^c$ and, for each $u \in \mathcal{U}^c$, treat the observation about u's mother as left truncated at $S_c^*(u)$.[33] Figure 3.11 shows the conditional survivor function calculated from the rates $\tilde{r}_c^*(t)$. This function obviously agrees quite well with the 1891-1900 period life table that was used to generate the data. Of course, the result would be basically the same if I had used a simple random sample from $\mathcal{U}^c$.

[33] One can use, therefore, any standard Kaplan-Meier procedure that allows for left truncated data. I have used TDA's `dple` procedure.

3.7.4. Retrospective Surveys

In the previous Section I assumed that I have data from a simple random sample from the complete set of children, $\mathcal{U}^c$. However, the data actually result from a retrospective survey performed in some specific year, say τ, and I therefore have to take into account that not all members of $\mathcal{U}^c$ survive until τ. Fortunately, the approach to estimate δ_t^f via the rates $\tilde{r}_c^*(t)$ that was discussed in the previous Section can also be applied to a retrospective sample if I make the additional assumption that children's life lengths are independent of their mother's life length.[34] To explain the argument, let T^c denote the life length of children in the reference set $\mathcal{U}^c$. On a historical time axis, if mothers are born in the year τ_0, each child $u \in \mathcal{U}^c$ survives until $\tau_0 + S_c^*(u) + T^c(u)$ (as already introduced, $S_c^*(u)$ is the age of the mother when u was born). The set of children who survive at least until the year τ is therefore given by

$$\mathcal{U}^c[\tau] := \{u \in \mathcal{U}^c \mid \tau_0 + S_c^*(u) + T^c(u) \geq \tau\}$$

In the simulation model introduced in Section 3.7.1 I assumed $\tau_0 = 1900$. Based on this assumption, Figure 3.6 shows the survival of children in historical time.

Now assume a retrospective survey performed in the year τ. The sample is then drawn from the reference set $\mathcal{U}^c[\tau]$. Following the approach discussed in the previous Section, I can calculate rates

$$\tilde{r}_{c,\tau}^*(t) := \frac{|\{u \in \mathcal{U}^c[\tau] \mid T_c^*(u) = t, S_c^*(u) \leq t\}|}{|\{u \in \mathcal{U}^c[\tau] \mid T_c^*(u) \geq t, S_c^*(u) \leq t\}|}$$

which are defined analogously to the rates $\tilde{r}_c^*(t)$ introduced in (3.15). In order to see that the rates $\tilde{r}_{c,\tau}^*(t)$ are reasonable estimates of the rates $\tilde{r}_c^*(t)$, their definition might be written in the following way:

$$\tilde{r}_{c,\tau}^*(t) = \frac{|\{u \in \mathcal{U}^c \mid T_c^*(u) = t, S_c^*(u) \leq t, S_c^*(u) + T^c(u) \geq \tau - \tau_0\}|}{|\{u \in \mathcal{U}^c \mid T_c^*(u) \geq t, S_c^*(u) \leq t, S_c^*(u) + T^c(u) \geq \tau - \tau_0\}|}$$

[34] This assumption, already built into the simulation model in Section 3.7.1, is probably not completely true. However, for the moment I will base my argument on this assumption.

The further argument proceeds in terms of conditional frequencies. Using an abbreviated notation, one may write:

$$\begin{aligned}
\tilde{r}^*_{c,\tau}(t) &= \frac{\Pr(T^*_c = t, S^*_c \leq t, S^*_c + T^c \geq \tau - \tau_0)}{\Pr(T^*_c \geq t, S^*_c \leq t, S^*_c + T^c \geq \tau - \tau_0)} \\
&= \frac{\Pr(S^*_c + T^c \geq \tau - \tau_0 \,|\, T^*_c = t, S^*_c \leq t)}{\Pr(S^*_c + T^c \geq \tau - \tau_0 \,|\, T^*_c \geq t, S^*_c \leq t)} \\
&\quad \times \frac{\Pr(T^*_c = t, S^*_c \leq t)}{\Pr(T^*_c \geq t, S^*_c \leq t)} \\
&= \tilde{r}^*_c(t)\, \frac{\Pr(S^*_c + T^c \geq \tau - \tau_0 \,|\, T^*_c = t, S^*_c \leq t)}{\Pr(S^*_c + T^c \geq \tau - \tau_0 \,|\, T^*_c \geq t, S^*_c \leq t)}
\end{aligned}$$

Now, given the assumption mentioned at the beginning, that, conditional on $S^*_c \leq t$, the survival of children does not depend on the survival of their mothers, the last term on the right-hand side becomes approximately

$$\frac{\Pr(S^*_c + T^c \geq \tau - \tau_0 \,|\, S^*_c \leq t)}{\Pr(S^*_c + T^c \geq \tau - \tau_0 \,|\, S^*_c \leq t)}$$

and may be omitted.

There is, however, a further difficulty resulting from retrospective surveys. The later the year τ in which the survey is performed, the smaller is the number of children who might participate in the survey, and consequently also the risk set to be used for the estimation of the death rates $\tilde{r}^*_{c,\tau}$ becomes smaller. This is shown in Figure 3.12 which is based on the data from the simulation model. Shown are the functions

$$t \longrightarrow \mathcal{U}^c_t[\tau] := \{u \in \mathcal{U}^c[\tau] \,|\, T^*_c(u) \geq t, S^*_c(u) \leq t\}$$

as they result from four fictitious retrospective surveys performed in the years $\tau = 2000, 2010, 2015$, and 2020. The possible problem concerns estimation with left truncated data. Contrary to the standard Kaplan-Meier procedure with right censored data only, the risk set is very small at the beginning and may not allow reliable estimates of the death rates. Due to the cumulative nature of the calculation of survivor functions from these rates, any imprecisions introduced at the beginning will then propagate to values of the survivor function at later ages. To illustrate,

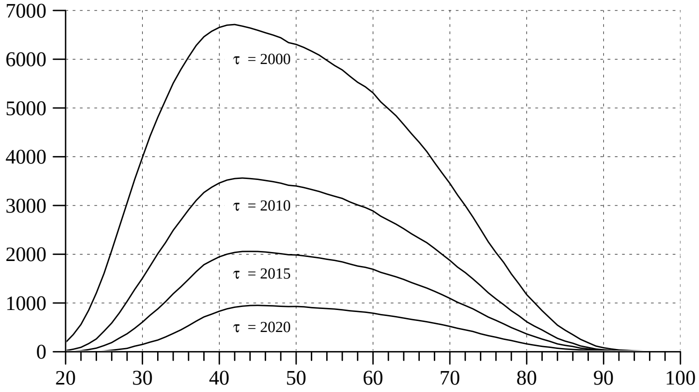


Figure 3.12.: Sizes of the risk sets $\mathcal{U}_t^c[\tau]$, depending on τ, calculated from four retrospective surveys of the simulated data set in the years $\tau = 2000, 2010, 2015$, and 2020.

I use the simulated data set and perform a retrospective survey in the year $\tau = 2010$. I assume that all children who survive this year, that is about 20% of the 11407 children in $\mathcal{U}^c$, participate in the survey and provide information about their mothers. Nevertheless, I can only begin to estimate a conditional survivor function at $a^+ = 25$ as shown in Figure 3.13.

3.8. Inferences from the GLHS and SOEP Data

I now use the methods discussed in the previous Sections to draw some inferences from the GLHS and SOEP data.

The *German Life History Study* (GLHS) is a long-term project conducted by the Max Planck Institute for Human Development (Berlin). The main data source of this project is a series of retrospective surveys in which members of selected birth cohorts were asked to provide detailed information about their life courses. Part of these data are accessible for the general scientific public through the *Zentralarchiv für empirische*

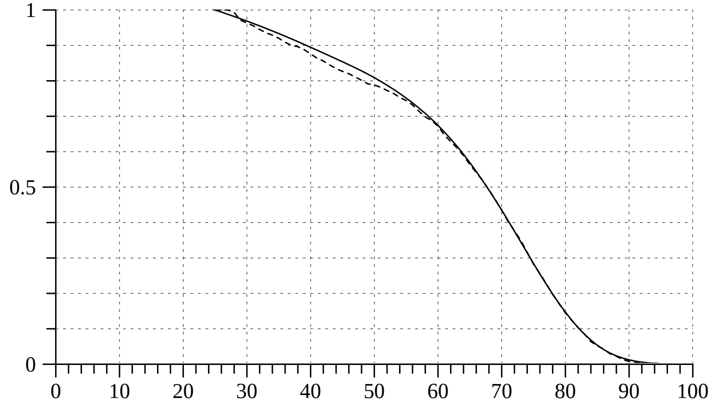


Figure 3.13.: Conditional survivor functions, estimated from $\mathcal{U}^c[2010]$ (dotted line) and calculated from the 1891–1900 period life table (solid line), both beginning at $a^+ = 25$.

Sozialforschung (Köln). Only the publicly accessible data are used in the following.

a) Data from the first survey (LV I) were sampled during the years 1981 – 83 and included 2171 members of the birth cohorts 1929 – 31, 1939 – 41, and 1949 – 51.

b) Data from a second survey (LV II) were sampled in two parts, both relating to persons born in the years 1919 – 21; a first part was conducted in 1985 – 86 and included 407 persons (LV IIA), a second part was conducted in 1987 – 88 and included 1005 persons (LV IIT).

c) Data from a third survey (LV III) were sampled in 1989 and included 2008 members of the birth cohorts 1954 – 56 and 1959 – 61.

All surveys were conducted in the territory of the former FRG. For the present study I include all female respondents from the surveys LV I, LV IIT, and LV III (only cohort 1959 – 61) having a German citizenship.

The number of cases and their distribution across the five cohorts is shown in the following table:[35]

Birth cohort	Birth years	Male	Female	Interview date
C20	1919 – 21	373	632	1987 – 88
C30	1929 – 31	349	359	1981 – 83
C40	1939 – 41	375	355	1981 – 83
C50	1949 – 51	365	368	1981 – 83
C60	1959 – 61	512	489	1989

3.8.1. Description of the Data

Of the 2171 respondents interviewed in LV I, 2120 respondents were able to provide a valid birth year of their mother. Of these mothers, 732 died before the interview date, 1386 were still alive, and for two mothers I have no information. Complete information is therefore available for 2118 mothers. In 8 cases this information is inconsistent or implausible, for example, the birth year of the respondent is greater than the death year of the mother. In addition to inconsistent cases I also exclude cases with a life length greater than 105 years. For women I also require that the age at which the women gave birth to her child (the respondent) is not greater than 51 years. If these cases are excluded, there finally remain 2110 cases in which the birth year of the mother, whether she died before the interview date, and, if she died, also her death year are known. Similarly, I get valid information for 2044 fathers.

The second study, LV II was conducted in two parts: LV IIA with interviews during 1985 – 86, and LV IIT with interviews during 1987 – 88. In the same way as explained for LV I I get valid information about the lifetimes of 387 + 956 = 1343 mothers and 382 + 943 = 1325 fathers. The third study, LV III, provides valid information about 1954 mothers and 1911 fathers.

[35] Of the 632 women of birth cohort C20 three did not give valid birth years for their children and will be excluded in further calculations.

Table 3.5.: Information about lifetimes of mothers and fathers available in the GLHS and SOEP data sets.

	LV I	LV IIA	LV IIT	LV III	SOEP
Interview dates	1981-83	1985-86	1987-88	1989	1986
Respondents	2171	407	1005	2008	8021
Mothers					
- valid birth year	2120	390	962	1954	7819
- still alive	1386	24	43	1766	4872
- known death year	732	366	919	188	2911
- no information	2	0	0	0	36
- complete information	2118	390	962	1954	7783
- dismissed	8	3	6	0	37
- remaining cases	2110	387	956	1954	7746
- still alive	1385	24	43	1766	4854
- died	725	363	913	188	2892
Fathers					
- valid birth year	2062	386	955	1916	7699
- still alive	909	1	8	1460	3586
- known death year	1150	384	945	451	4053
- no information	3	1	2	5	60
- complete information	2059	385	953	1911	7639
- dismissed	15	3	10	0	25
- remaining cases	2044	382	943	1911	7614
- still alive	909	1	8	1460	3577
- died	1135	381	935	451	4037

Comparable information is available from the third wave of the *SOEP* conducted in 1986. All members of subsample A of the SOEP were asked to provide information about birth years of their parents, whether parents died before the interview date and, if they died, about their death years. In order to get data comparable with the GLHS, I selected only persons with a German citizenship. As shown in Table 3.5, there are 8021 persons providing valid information about 7746 mothers and 7614 fathers. Taking the GLHS and SOEP data together, I finally have valid information about 13153 mothers and 12894 fathers.

I prepared two data files for further analysis, one for mothers and the other one for fathers. Both files contain values of four variables:

B^f := birth year of the mother

P^f := birth year of the child (respondent)

E^f := 1 if mother died before the interview date, 0 otherwise

D^f := mother's death year, or the year of the interview, depending on the value of E^f

Variables in the data file for fathers are defined accordingly and will be denoted by B^m, P^m, E^m, and D^m.

3.8.2. Survivor Functions of Parents

I now apply the method discussed in the previous Section to the data introduced in Section 3.8.1. Since mortality conditions have substantially changed during the last 100 years, I consider birth cohorts as defined in Table 3.6.[36] To develop the argument I consider variables $\hat{T}_c^f$ and $\hat{T}_c^m$ representing the life length of women and men who belong to a birth cohort indexed by c. Derivable from the variables introduced at the end of Section 3.8.1, available data are given by variables

$$C_c^f := P_c^f - B_c^f \quad \text{and} \quad C_c^m := P_c^m - B_c^m$$

which record the ages at which persons belonging to birth cohort c became mothers or fathers, and variables

$$T_c^f := D_c^f - B_c^f \quad \text{and} \quad T_c^m := D_c^m - B_c^m$$

which record the knowledge about the life length. If $E_c^f(u) = 1$, $T_c^f(u) = \hat{T}_c^f(u)$ is the known life length of u; otherwise, the information is censored and I only know that $\hat{T}_c^f(u) \geq T_c^f(u)$. For variables pertaining to men the interpretation is analogous.

[36] Compared with the figures in Table 3.5 the total number of cases is slightly smaller because persons born before 1870 or after 1939 have been omitted.

Table 3.6.: Definition of birth cohorts used in the estimation of survivor functions.

		Mothers			Fathers		
Cohort	Birth years	died	alive	total	died	alive	total
C1	1870 – 1879	271	0	271	528	0	528
C2	1880 – 1889	1064	10	1074	1393	12	1405
C3	1890 – 1899	1698	170	1868	1591	101	1692
C4	1900 – 1909	1123	954	2077	1685	600	2285
C5	1910 – 1919	438	1456	1894	907	1011	1918
C6	1920 – 1929	272	2467	2739	464	2035	2499
C7	1930 – 1939	123	2219	2342	196	1773	1969

t	(a)	(b)	(c)	(d)
15	1	0	0	1.000
16	2	0	0	1.000
17	3	0	0	1.000
18	18	0	0	1.000
19	45	0	0	1.000
20	100	1	0	1.000
21	183	1	0	0.990
22	266	2	0	0.985
23	347	0	0	0.977
24	435	2	0	0.977
25	556	1	0	0.973
26	690	3	0	0.971
27	781	4	0	0.967
28	928	1	0	0.962
29	1075	3	0	0.961
30	1217	3	0	0.958
31	1327	2	0	0.956
32	1427	5	0	0.954
33	1512	9	0	0.951
34	1597	5	0	0.945
35	1677	6	0	0.942
36	1740	5	0	0.939
37	1800	5	0	0.936
38	1864	9	0	0.934
39	1903	13	0	0.929
40	1929	12	0	0.923
41	1945	2	0	0.917
42	1952	5	0	0.916
52	1901	4	0	0.878
53	1897	15	0	0.876
54	1882	13	0	0.869
55	1869	21	0	0.863
56	1848	7	0	0.854
57	1841	17	0	0.850
58	1824	14	0	0.843
59	1810	18	0	0.836
60	1792	20	0	0.828
61	1772	19	0	0.818
62	1753	15	0	0.810
63	1738	23	0	0.803
64	1715	18	0	0.792
65	1697	33	0	0.784
66	1664	23	0	0.769
67	1641	36	0	0.758
68	1605	32	0	0.741
69	1573	24	0	0.727
70	1549	41	0	0.715
71	1508	39	0	0.697
72	1469	54	32	0.679
73	1383	53	53	0.654
74	1277	56	57	0.629
75	1164	53	43	0.601
76	1068	55	30	0.574
77	983	42	125	0.544
78	816	57	96	0.521
79	663	42	114	0.484

t	(a)	(b)	(c)	(d)
43	1957	7	0	0.914
44	1957	5	0	0.910
45	1956	13	0	0.908
46	1947	7	0	0.902
47	1943	10	0	0.899
48	1935	8	0	0.894
49	1928	8	0	0.891
50	1920	7	0	0.887
51	1913	12	0	0.884
80	507	33	85	0.454
81	389	20	78	0.424
82	291	12	70	0.402
83	209	21	48	0.386
84	140	12	34	0.347
85	94	5	41	0.317
86	48	0	39	0.300
87	9	0	5	0.300
88	4	0	4	0.300

Table 3.7.: Mortality data for mothers belonging to birth cohort C4 in the merged GLHS and SOEP data set. (a) Size of risk set at age t. (b) Number of deaths at age t. (c) Number of censored cases at age t. (d) Values of the conditional survivor function at age t.

To illustrate the calculations I refer to women belonging to birth cohort C4. The data are shown in Table 3.7. The column labelled (a) shows the risk sets. As discussed in the previous Section, the risk set at age t contains all women who did not die before t and became a mother not later than t.[37] In this example, the youngest age for which I know of a child is 15; risk sets can therefore be calculated only for ages $t \geq t^* = 15$. The next column, labelled (b), shows the number of death events. Then follows column (d) providing the number of censored cases which are required to update the risk sets. As shown by the definition

$$\mathcal{R}^*(t) := \{u \mid T_c^f(u) \geq t, C_c^f(u) \leq t\}$$

women belong to a risk set only until the maximal value of T_c^f, that is, until a death event occurs or until the interview date (of their children).

The information in Table 3.7 suffices to calculate death rates. For example, $r^*(20) = 1/100$ and $r^*(80) = 33/507$. These rates can then be used to

[37] Of course, from the data I do not know when women actually gave birth to a first child. Whether this has implications for the quality of the estimates will be discussed in a later Section.

estimate the survivor function

$$G^*(t) = g_{t^*} \prod_{j=t^*}^{t-1} (1 - r^*(j))$$

Of course, I do not know g_{t^*}, that is, the proportion of women who survived age 14. So I can only estimate a conditional survivor function

$$G(T_c^f \mid T_c^f \geq t^*) \approx \prod_{j=t^*}^{t-1} (1 - r^*(j))$$

which is shown in the last column of Table 3.7 labelled (d).

Since this approach to estimate a conditional survivor function depends on a previous estimation of rates, one should also consider the question whether these rates can be reliably estimated. Formally, one can begin at age t^* which is 15 in the example. However, due to the small number of cases in the risk sets at ages under 20, one might question the reliability of these estimates. In fact, the estimation procedure implies estimated death rates having a value of zero during ages from 15 to 19. But given our knowledge about mortality and life tables from other sources, these estimates will clearly be wrong. Moreover, the reliability of estimates of death rates not only depends on the size of the risk sets but also on the number of death events that can be observed. With regard to the data in Table 3.7, it might be better to begin an interpretation of estimated death rates only at some later age, for example, at age 26 or even later.

Conditional survivor functions can be represented graphically in two possible ways: The function can be plotted beginning at some age t with arbitrary value g_t; or one can try to find some estimate of g_t and then plot the conditional survivor function as part of a complete survivor function. In any case one needs to decide where to start the plotting. For the example I begin at age 26 and estimate g_{26} from the female survivor function of the German period life table for the period 1901–10 (see Table 3.3). Beginning at age 26, I therefore multiply all values of column (d) in Table 3.7 with the factor $g_{26} = 0.71463/0.971 = 0.736$. The result is shown in Figure 3.14. The dotted line represents the female survivor

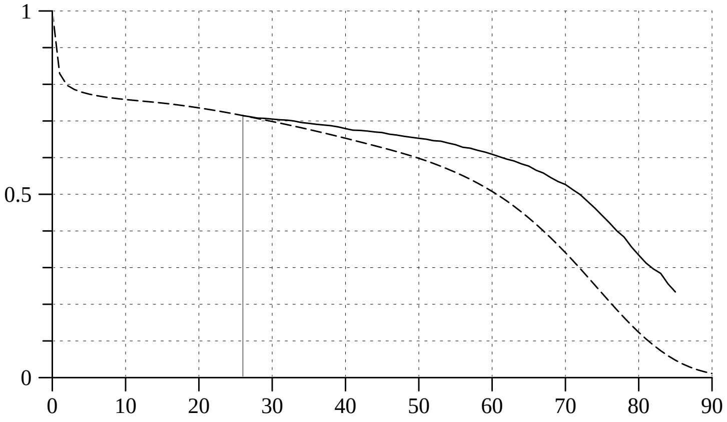


Figure 3.14.: Female survivor function of the German period life table 1901/10 (dotted line) and conditional survivor function from Table 3.7.

function from the 1901–10 period life table; the solid line shows the adjusted conditional survivor function from Table 3.7. By definition, values are identical at age 26. The different development of both curves reflects the reduction of death rates that occurred during the period from about 1930 until the end of the century. So I might use the latest 1986 –88 period life table for a further comparison. As can be estimated from Table 3.7, the death rate at age 80 is about 0.065. A corresponding estimate from the 1986 –88 period life table is 0.066.[38] One should note, however, that values of rates calculated from sample data for single years often show high fluctuations and it might be better, therefore, to use smoothed values based on larger age classes.

In the same way as has been discussed for women belonging to birth cohort C4 (1900 –1909) one can estimate conditional survivor functions for all birth cohorts in Table 3.6. Results are shown in Figure 3.15. To allow for a comparison, all survivor functions are drawn conditional on $t^* = 30$. The placement onto a historical time axis was done by using the centres of the birth cohort intervals. For example, the value of the

[38] Calculated from the data in Table 3.3.

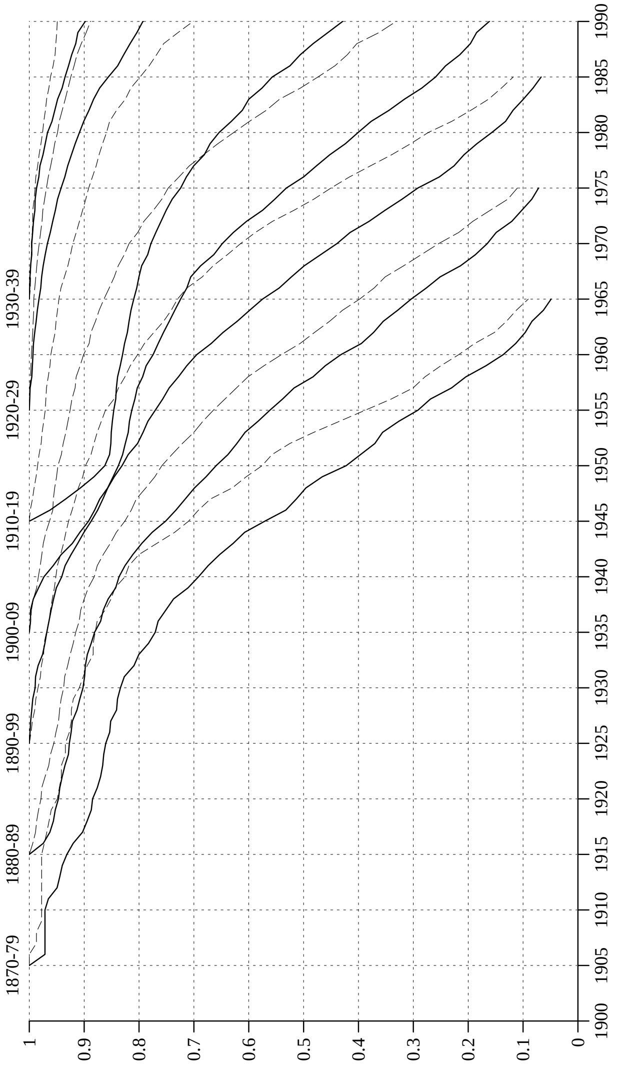


Figure 3.15.: Conditional survivor functions, beginning at age 30, for men (solid lines) and women (dotted lines) belonging to specified birth cohorts.

conditional survivor function for birth cohort C1 at age 30 is shown in the year 1875 + 30 = 1905. The changing shapes of the survivor functions not only reflect a general tendency of decreasing death rates, both for men and women. Also clearly seen are period effects, especially the substantial increases of male death rates during the years of World War II. This seems not to be the case with regard to female death rates. An interpretation should consider, however, that the occurrence of death events might not be independent for mothers and their children, in particular during war time. The death events of mothers might therefore be substantially underrepresented in the data set.

3.8.3. Visualisation of Death Rates

In order to investigate period effects it is often preferable to directly plot the rates from which (conditional) survivor functions are derived. The only drawback is that rates calculated from small samples are often highly fluctuating. As an example I refer to death rates of men belonging to birth cohort C5 (1910 –1919). The solid line in Figure 3.16 shows the death rates as directly calculated from the data, that is, for each year of age, the number of deaths divided by the number of persons in the risk set. There obviously are big fluctuations. One should therefore apply some kind of smoothing procedure to provide a better view of the general shape of the rate function.

Many such smoothing procedures have been proposed in the literature. In the present context, smoothing will only serve to visualise rate functions. It might therefore suffice to simply use moving averages. Given a series of values r_t, for $t = t_1, \dots, t_n$, each value is then substituted by a mean of neighbouring values. If the number of neighbours is denoted by $2k$, the smoothed values are calculated as

$$r_t^{(k)} := \frac{1}{2k+1} \sum_{j=t-k}^{t+k} r_j$$

At both ends of the series only the actually available values are taken into

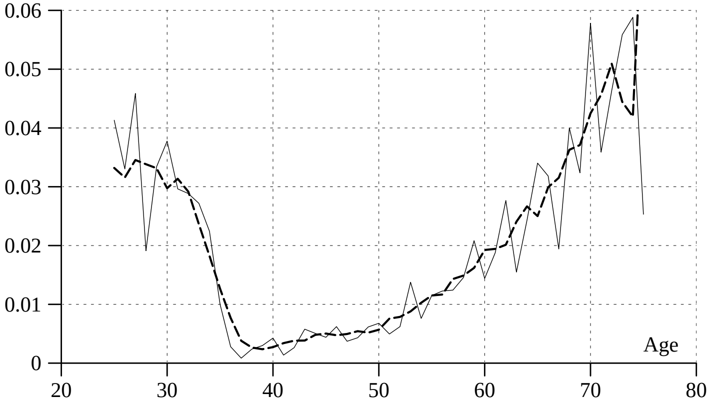


Figure 3.16.: Raw values (solid line) and smoothed values (dotted line) of death rates of men belonging to birth cohort C5 (1910 – 1919).

account. The complete formula may then be written as follows:

$$r_t^{(k)} := \frac{1}{\min\{t_n, t+k\} - \max\{t_1, t-k\} + 1} \sum_{j=\max\{t_1,t-k\}}^{\min\{t_n,t+k\}} r_j$$

where t_1 and t_n refer, respectively, to the first and last element of the series.

Choosing $k = 2$, this procedure was used to calculate values for the dotted line in Figure 3.16. It is seen how the smoothing removes the fluctuations but preserves the global shape of the rate function.

I now compare the death rates of men belonging to birth cohorts C1, . . . , C6. The rate functions are shown in Figure 3.17 and placed onto a historical time axis. To support visibility, the rate functions are smoothed with the procedure just described (again, $k = 2$). Compared with the survivor functions shown in Figure 3.15, the rate functions provide a much better view of the impact of World War II.

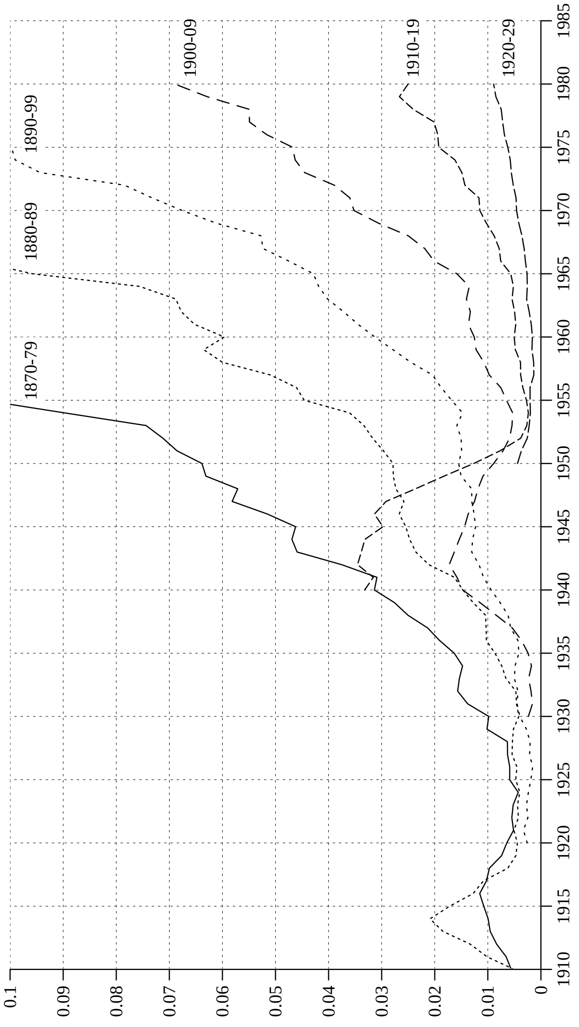


Figure 3.17.: Smoothed death rates of men belonging to the indicated birth cohorts. (Moving averages with $k = 2$.)

3.9. Conclusion

The discussion in Section 3.7 and 3.8 was based on a rather simplistic simulation. The estimators derived from relative frequency analogues of dynamic locally dependence conditions discussed in Section 3.6 could be applied successfully within this context. The dynamic independence conditions is necessarily formulated within the context of probability models that provide the connection between the 'full data' and 'left truncated data'. In fact, the dynamic independence condition was explicitly employed in the probabilistic description of the simulation algorithm used to gauge the performance of suggested estimators. It is therefore not surprising to find good agreement between the period survival rates and estimates based on the simulated birth and death events.

The fact that survival rates and their empirical analogues can be defined without recourse to the number of children born to a woman (see Section 3.7.3) illustrates the difference from a pure sampling based argument as developed in Section 3.3. While the latter does not depend on any form of probabilistic independence assumption and the correctness of its computations can be checked against empirical findings, Section 3.4 has shown that this will be impossible for the dynamic local independence condition incorporated in the construction of weights and estimators. An argument for the correctness of the dynamic independence conditions must necessarily refer to a probability model of the dynamic population process. It presupposes the use of a probability model. Therefore, the basic ingredients of any dynamic probability model are a necessary part of a constitutive decision in Matheron's sense, the decision to model population dynamics in terms of dynamic probabilistic models. The appropriateness of the dynamic local independence assumption can only be judged within such a framework.

But the model used in the simulation is a rather simplistic one and misses one important ingredient inherent in any study of real populations: The simulation created sets $\mathcal{U}$ of potential mothers "“born"” at a fixed point in time, without reference to their parent generation. The arguments in Section 3.7 then showed how the local independence condition allows the development of reasonable estimators for such constructed cohorts.

In the applications, the reasoning was applied to cohorts artificially constructed from retrospective information. But real populations grow or shrink depending on both birth and death rates of previous generations. Even when ignoring possible dependencies between life length of parents and their children, the procedure can only be justified within a probability model that posits a simple random process creating parents. This has been proposed by Brillinger (1986) who tried to justify a Poisson process as a model of creating parents. A crucial consequence of such a proposal was pointed out by Hoem in the discussion of Brillinger's paper.

> Unfortunately, the assumption of Poisson births disregards essential aspects of the internal dynamics of real-life populations. Results based on the Poisson assumption must have little relevance for the analysis of statistics which involve data for several generations ... In normal populations, children are borne by population members ... Since real populations are subject to the internal dependencies which are made manifest by population waves, the assumption of Poisson births is untenable in multigenerational analyses.

The problem surfaces also in our problem of estimating parent's length of life: While the number and and ages at birth of a mother are of no direct importance to the question posed, they necessarily appear in any analysis of the probability model of parent's life length. It is this direct dependence of current populations on the timing and number of births of previous generations that makes the Poisson models untenable. At the same time, number of births and their timing are closely related to life length. Thus, they ought to be part of the probabilistic models. But their mere presence changes the form of reasonable estimators.[39]

But reasonable probabilistic models are very difficult to construct since it turns out that Markov structures on a calender time basis are inappropriate and intergenerational dependencies invalidate transition models

[39] In their discussion of the construction of likelihood functions, Bayarri et al. (1987) and Bayarri and DeGroot (1992) use selection models of the present type to demonstrate that various suggestions as to what ought to figure in a likelihood function are ill defined.

based on individuals alone.[40]

If and in so far such an elaborated probabilistic model of population dynamics can be combined with dynamic local independence conditions, such models do satisfy the requirements for scientifically valuable stochastic models in the sense of Neyman. The repeated application of local dynamic independence conditions does lead to manageable estimators and manageable estimation strategies. But on its own the independence condition does not provide arguments for its support. Criticism of this condition must be based on a larger probability model of population dynamics as proposed by Jagers (1989) or similar ones.

[40] Jagers (1989) proposed a clever construction that allows to keep a Markov assumption for lines of descent. Life length, number of children, and other properties of interest can then be modelled as attributes of elements of these lines of descent. He himself describes the purpose of such a construction as "‘But one of the purposes of mathematical population dynamics is to serve as a bridge between individual properties and properties of the population as a whole. …This presupposes a theory of population evolution built upon a description of individual life. But individuals vary …, i.e. we need a stochastic description."’ Jagers (1989). My interest here is of course a different one.

4

Probabilistische Selektionsmodelle

4.1. Summary

Missing or incomplete data are a pertinent to any social survey. Interviewees do not answer some questions or refuse to participate at all. Nowadays, response rates in surveys rarely reach 40%. Can the answers of 40% of the selected individuals be treated as if one had answers from all of them? Probabilistic selection models can be used to speculate on the connexion between the attained information and the possible answers of the non-responders. These models can be extended to cover incomplete data such as grouped or censored data as well as processes of self-selection and the evaluation of social programmes. A brief review of some of these applications is given. Next, it is shown that the assumptions used in probabilistic selection models necessarily refer to concepts transcendenting empirically accessible social facts. However, the amount and impact of these assumptions can be reduced if possibly available additional incomplete information is fully used. Therefore, probabilistic selection models are formulated in a form that allows for grouped, censored and general coarsened data. Finally, several proposals for the sensitivity analysis of probabilistic selection models are presented and it is argued that approaches restricted to too small neighbourhoods of a given model can be quite misleading. Throughout, responses on

income questions obtained in the German ALLBUS 1996 are used as an illustration.

4.2. Zusamenfassung

Fehlende oder unvollständige Angaben in Umfragedaten sind unvermeidlich. Befragte geben manchmal keine Auskunft oder verweigern das Interview. Der Anteil der vollständigen Antworten ist selten größer als 40%. Was weiß man über die anderen 60%? Kann man problemlos die Angaben der 40% so behandeln als hätte man Angaben von allen ausgewählten Befragten? Probabilistische Selektionsmodelle sind ein Hilfsmittel, über Zusammenhänge zwischen erhaltenen und nicht erhaltenen Angaben nachzudenken. Ihre Anwendungsmöglichkeiten beschränken sich nicht auf die Analyse fehlender Angaben in Umfragen. Sie werden auch für die Untersuchung gruppierter, zensierter oder sonstiger unvollständiger Daten sowie für die Modellierung von Prozessen mit Selbstselektionen und die Evaluation von Maßnahmen benutzt. Nach einem kurzen Überblick über verschiedene Anwendungen wird zunächst diskutiert, welche Annahmen Selektionsmodellen zugrunde liegen. Es zeigt sich, dass die Verwendung probabilistischer Modelle auf Voraussetzungen verweist, die über empirisch zugängliche Annahmen über soziale Sachverhalte weit hinausgehen. Der Anteil modellimmanenter Spekulation kann allerdings verringert werden, wenn auch partielle Angaben der Befragten mit einbezogen werden. Die Grundlagen von Selektionsmodellen werden daher nicht nur für fehlende, sondern auch für gruppierte und vergröberte Angaben formuliert. Abschließend werden verschiedene Möglichkeiten der Sensitivitätsanalyse für probabilistische Selektionsmodelle dargestellt und die Beschränkungen enger Formulierungen demonstriert. Fehlende und gruppierte Einkommensangaben im ALLBUS 1996 dienen zur Illustration.

4.3. Einleitung

Fehlende oder unvollständige Angaben in Umfragedaten sind unvermeidlich. Schließlich steht es jedem Befragten frei, die Zumutung eines Interviews zurückzuweisen oder auf einzelne Fragen nicht zu antworten. Zudem mag ein Befragter nicht willens oder in der Lage sein, Auskünfte in der gewünschten Präzision zu geben. In Umfragen ist der Anteil vollständiger Angaben zu einer bestimmten Frage selten höher als 40%. Dennoch sollen Umfrageergebnisse Aufschluss über soziale Sachverhalte geben. Wenn dies nachvollziehbar gelingen soll, muss ein Zusammenhang zwischen den erhaltenen Angaben und den potentiellen Angaben aller ausgewählten Befragten hergestellt werden. Denn allein aufgrund der vollständigen Angaben soll es möglich sein, Aussagen über statistische Sachverhalte zu treffen, die sich auf alle Befragten beziehen. Wenn aber über 60% der Befragten nichts oder wenig bekannt ist, dann sind zur Rechtfertigung solcher Aussagen offenbar starke Annahmen erforderlich. Probabilistische Selektionsmodelle sind ein Hilfsmittel, solche Annahmen explizit zu formulieren und über ihre Konsequenzen nachzudenken.

Das Grundschema probabilistischer Selektionsmodelle lässt sich leicht angeben: Man möchte Aussagen über Aspekte der Verteilung einer statistischen Variablen Y machen. Nur wurden die Werte dieser Variablen nicht vollständig beobachtet. Dagegen ist die Verteilung einer Variablen Y^* bekannt, von der angenommen wird, sie stände in einer Beziehung zu der interessierenden Variablen Y. Zum Beispiel mag ein Befragter auf die Frage nach dem Einkommen mit einer genauen Angabe, mit einer ungefähren Angabe etwa in Form eines Intervalls, oder gar nicht antworten. Der Wert der Variablen Y^* ist dann entweder der Wert von Y, oder ein Intervall, in welches Y fällt, oder das Intervall $(0, \infty)$, das „keine Angabe“ repräsentiert. Wäre bekannt, unter welchen Umständen jemand bei gegebenem Wert von Y eine mehr oder weniger genaue Auskunft Y^* gibt, dann könnte aus der Verteilung von Y^* und den Umständen auf Aspekte der Verteilung von Y geschlossen werden. Wird für die Antwortmöglichkeiten Y^* bei gegebenem Y eine bedingte Wahrscheinlichkeit angegeben, dann soll dieses probabilistische Modell im Folgenden ein allgemeines *Selektionsmodell* für (Y, Y^*) heißen.

In der Literatur zur Stichprobentheorie sind probabilistische Selektionsmodelle nur sehr zurückhaltend diskutiert worden. So schrieb Tore Dalenius, damals Präsident der International Association of Survey Statisticians:

> I take a dim view of the usefulness of these endeavors on two grounds. (1) First, it appears utterly unrealistic to postulate 'response probabilities' which are independent of the varying circumstances under which an effort is made to elicit a response. …(2) …it seems unavoidable to introduce assumptions of unknown validity about probabilities. In summary, I am inclined to reject approaches to the nonresponse problem which involve 'response probabilities' (Dalenius in Madow und Olkin 1983, Band 3: 412).

Und Mohler et al. verweisen auf die großen Schwankungen in den Ausschöpfungsraten von Stichproben, „die sich nicht mehr auf statistischen Zufall zurückführen lassen" (Mohler et al. 2003: 11). Daher hat man sich in dieser Tradition darauf beschränkt, Bereiche möglicher Schlussfolgerungen auszuweisen, die sich allein auf die Angaben der Befragten und Konsistenzannahmen stützen.

Dagegen sind sowohl in der mathematischen Statistik wie auch in der Biometrie und Ökonometrie seit etwa 30 Jahren eine Fülle von probabilistischen Selektionsmodellen entwickelt worden. Einerseits ist geklärt worden, unter welchen Modellvorstellungen relativ einfache Verfahren des Umgangs mit fehlenden Daten gerechtfertigt werden können. In solchen Modellen können die Einzelheiten des Zustandekommens unvollständiger Daten weitgehend ignoriert werden. Einen guten Überblick darüber geben die Bücher von Schafer (1997) und Little und Rubin (2002). Andererseits sind in der ökonometrischen Tradition hauptsächlich nicht ignorierbare Selektionsmodelle und entsprechende Schätzverfahren vorgeschlagen worden. Insbesondere die frühen Arbeiten von Heckman (1976, 1979) haben einen kaum zu überschätzenden Einfluss auf viele Bereiche der Sozialwissenschaften ausgeübt. Neuere Überblicke geben Nicoletti (2002) und Vella (1998). Obwohl die Abhängigkeit dieser Modelle von einer unübersichtlichen Mischung von Annahmen über Verteilungen, funktionale Formen von Regressionen, latente Variablen und Ausschlussrestriktionen früh kritisiert wurde (z.B. in Wainer 1986, 1989), haben sie sich in einigen Bereichen der Sozialwissenschaften als dominante analytische Methode durchgesetzt.

Dieser Artikel gibt einen Überblick über neuere Entwicklungen in der statistischen Literatur. Insbesondere werden die Form von Annahmen, die Einbeziehung unvollständiger Angaben und die Verwendung von Sensitivitätsanalysen diskutiert. Im nächsten Abschnitt wird zunächst ein kurzer Überblick über häufig verwandte probabilistische Selektionsmodelle und ihre Anwendungen in den Sozialwissenschaften gegeben. Anschließend wird am Beispiel der Einkommensangaben im ALLBUS 1996 die Form der Annahmen diskutiert, die in probabilistische Modelle einfließen. Die Annahmen verweisen nicht nur auf potentiell empirisch zugängliche soziale Sachverhalte, sondern enthalten immer auch modellimmanente Anteile, die nicht reifiziert werden sollten. Das mag die Zurückhaltung gegenüber solchen Modellen in der Stichprobentheorie rechtfertigen. Allerdings kann der Anteil modellimmanenter Spekulation verringert werden, wenn auch gruppierte, zensierte und andere partielle Angaben in die Analyse einbezogen werden. Abschnitt 4.7 beschreibt die Vorgehensweise für den Fall ignorierbarer Selektionsmodelle. Im Abschnitt 4.8 wird die Ignorierbarkeit von Selektionsmodellen auch für gruppierte und vergröberte Angaben in einem wahrscheinlichkeitstheoretischen Rahmen definiert. Anschließend werden zwei Schätzverfahren vorgestellt, die die Einbeziehung von Kovariablen in Selektionsmodelle erlauben. Einige nicht ignorierbare Modelle werden im folgenden Abschnitt kurz vorgestellt. Abschließend werden Techniken der Sensitivitätsanalyse dargestellt. Sie erlauben eine Abschätzung der Abhängigkeit von Schlussfolgerungen von einigen zentralen modellimmanenten Annahmen und sind daher ein wesentliches Hilfsmittel für die Beurteilung von Selektionsmodellen.

Wenn möglichst alle stochastischen Annahmen eines Selektionsmodells systematisch variiert werden, so zeigt sich, dass der Bereich möglicher Schlussfolgerungen probabilistischer Modelle sehr groß werden kann. Zudem deckt er sich häufig mit den Bereichen, die im Rahmen der klassischen Stichprobentheorie entwickelt wurden. Globale Sensitivitätsanalysen probabilistischer Selektionsmodelle führen daher zu Abschätzungen, die mit denen der Stichprobentheorie vergleichbar sind. Die Formulierung verschiedener Annahmen in probabilistischen Selektionsmodellen ermöglicht eine Diskussion über den Zusammenhang zwischen erhaltenen und nicht erhaltenen Angaben, die stichprobentheoretische Überlegungen ergänzen können.

4.4. Selektionsmodelle in den Sozialwissenschaften

Umfragedaten bilden eine wesentliche empirische Grundlage aller Sozialwissenschaften. Aber schon das Verfahren, mit dem Befragte ausgewählt werden, ist häufig mit dem Hinweis hinterfragt worden, die Auswahl sei selektiv gewesen. Ein oft und gern zitiertes Beispiel ist der spektakuläre Misserfolg der Wahlvorhersage der Zeitschrift Literary Digest für die Präsidentenwahl 1936 in den USA. Das Literary Digest hatte 60% der Stimmen für den Republikaner Landon vorhergesagt, aber Roosevelt gewann die Wahl mit 62%. Das Literary Digest hatte eine Stichprobe von Telefon- und Autobesitzern befragt. Eine klassische Erklärung des Fehlschlags besagt, Telefon- bzw. Autobesitz sei damals ein Anzeichen von Reichtum gewesen und reichere Personen hätten eher republikanisch gestimmt. Das Selektionsargument bezieht sich auf den gewählten Rahmen der Stichprobe, die Basis für die Auswahl von Befragten. Aber eine einfache Überlegung zeigt, dass diese Selektion nicht allein für den Misserfolg verantwortlich sein kann. 1936 besaßen ca. 40% der Haushalte ein Telefon. Hätten die Telefon- und Autobesitzer in der Tat zu 60% für Landon gestimmt, dann hätten von allen Haushalten, die weder Auto noch Telefon besaßen, über 75% für Roosevelt stimmen müssen. Betrachtet man die abgegebenen Wählerstimmen, so hätte der entsprechende Anteil sogar größer als 90% sein müssen (Bryson 1976). Das Verhältnis der Odds für Roosevelt in den beiden Gruppen der Telefonbesitzer und derjenigen ohne Telefon müsste also mehr als 1:20 betragen. Das Literary Digest hatte 10 Millionen Fragebogen verschickt, aber nur 2.3 Millionen zurückerhalten. Es liegt nahe zu vermuten, dass Landon-Anhänger eher als Roosevelt-Anhänger auf die Umfrage geantwortet haben. Wird angenommen, die 10 Millionen Befragten hätten tatsächlich zu 62% für Roosevelt gestimmt, muss das Verhältnis der

Odds für Roosevelt in den beiden Gruppen der Antwortenden und der nicht Antwortenden nur 1:3 betragen, um zu der Diskrepanz zwischen Vorhersage und Wahlergebnis zu führen. Ein Verhältnis der Odds von 1:3 ist eine weit realistischere Größenordnung als die 1:20, die für die These einer Selektion zwischen Telefonbesitzern und Nichtbesitzern angenommen werden müsste. Daten aus Nachbefragungen haben dann auch die Bedeutung der Unterschiede im Antwortwortverhalten der Landon- bzw. Rooseveltanhänger bestätigt (Squire 1988; Cahalan 1989). Squire (1988: 132) schließt:

> The analysis here should also call attention to the other potential problem with any survey: nonresponse bias. ... Consumers of public opinion surveys, as well as practitioners, must be reminded of this potential problem in order to avoid a future disaster like the *Literary Digest* poll of 1936.

Brysons Überlegungen über den Misserfolg der Wahlvorhersage des Literary Digest verweisen zwar auf die möglicherweise gravierenden Folgen unvollständiger Angaben in Umfragen, benutzen aber keine probabilistischen Modelle etwa über das Antwortverhalten von Landon- bzw. Rooseveltanhängern. Seit der Mitte der 70er Jahre sind Verfahren entwickelt worden, die auf der Basis probabilistischer Modelle für Teilnahme- und Antwortentscheidungen der Befragten versuchen, die Folgen unvollständiger Angaben abzuschätzen. Eine Variante, die auf Arbeiten von Heckman (1976, 1979, 1990) zurückgeht, ist von Engelhardt (1999) vorgestellt worden. Sie untersucht die Einkommensangaben in der Berliner Altersstudie (BASE), einer nach Alter und Geschlecht geschichteten Zufallsstichprobe von Berlinern und Berlinerinnen über 70 Jahren, die auf der Basis des Einwohnermelderegisters gezogen wurde (Mayer und Baltes 1996). An der Erstbefragung haben 928 Personen teilgenommen, das entspricht ca. 49% der Ausgangsstichprobe. Engelhardt verwendet neben dem Einkommen die Variablen Geschlecht, Alter, Familienstand, Schul- und Berufsausbildung, Wohnform, Interviewform sowie einen Demenzindikator. Vollständige Angaben zu diesen Variablen liegen für 842 Personen vor, zusätzliche Angaben zum Einkommen nur für 716 oder 77% der Personen, die an der Befragung teilgenommen haben. Engelhardt (1999: 716f) unterstellt eine Wahrscheinlichkeit für die

Antwort jeder Person auf die Einkommensfrage, die über einen Probit-Link linear von allen Variablen (bis auf den Familienstand) abhängt. Außerdem nimmt sie an, das logarithmierte Einkommen aller Personen habe eine lineare, homoskedastische Regression auf alle Variablen (bis auf Alter, Demenzindikator und Interviewform) und folge einer bedingten Normalverteilung. Unter diesen Annahmen kann aus der bedingten Verteilung der beobachteten Einkommen auf die bedingte Verteilung der Einkommen aller Personen geschlossen werden. Engelhardt vergleicht die entsprechenden Ergebnisse mit einer linearen Regression, die nur die vollständigen Angaben berücksichtigt. Sie schließt,

> die Heckman-Korrektur [bietet] aber auch Möglichkeiten, die in der explorativen Analyse liegen. Wenn—wie im Beispiel—die selektionskorrigierte Regressionsanalyse zu demselben Resultat kommt wie die unkorrigierte Schätzung, erhöht dies das Vertrauen in die OLS-Regression (1999: 721).

Sie betont aber die Abhängigkeit der Ergebnisse von den unterstellten Annahmen (1999: 713f) und zeigt, dass zumindest diejenigen Annahmen, die sich überprüfen lassen, wohl nicht gelten (1999: 719). Um tatsächlich ein erhöhtes „Vertrauen in die unkorrigierte Schätzung“ zu haben, müsste an Stelle eines einzigen alternativen Selektionsmodell, das zudem auf zweifelhaften Annahmen beruht, mehrere Selektionsmodelle verglichen werden. Heckmans Modell bietet aber keinen systematischen Ansatzpunkt, Modellannahmen zu variieren bzw. die Auswirkungen verletzter Annahmen quantitativ abzuschätzen. Die Möglichkeit, mit Hilfe von probabilistischen Selektionsmodellen über die Folgen unvollständiger Angaben nachzudenken, kann im Rahmen dieses Modells nur begrenzt genutzt werden.

In anderen Bereichen, etwa der historischen Demographie, werden dagegen manchmal probabilistische Selektionsmodelle eingesetzt, um über die Aussagekraft von Angaben zu spekulieren. So existieren oft nur unvollständige Angaben über die Lebensdauer von Menschen. Z.B. gibt es zu Geburts- bzw. Taufangaben in Kirchenregistern in vielen Fällen keine Angaben über das Todesdatum. Allerdings gibt es manchmal weitere Ereignisse wie Heiraten oder Kindergeburten, die aufgezeichnet wurden. Es folgt, dass die Person mindestens das Alter bei diesem Ereignis erreicht hat. Ist Y das Lebensalter einer Person, dann ist Y^* entweder der Wert von Y, falls ein Todesdatum registriert wurde, oder aber das Intervall $(T_{\max}, \infty)$, wobei $T_{\max}$ der Zeitpunkt des letzten Ereignisses ist, das registriert wurde. Wenn angenommen wird, diese Angaben sagten nichts anderes als das jemand älter als $T_{\max}$ geworden ist, dann können Verfahren der Ereignisanalyse eingesetzt werden. Dagegen kann eingewandt werden, der Grund des Fehlens eines Todesdatums sei i.d.R. die Abwanderung der Personen. Dann wäre das Datum der Abwanderung eine untere Grenze für das erreichte Lebensalter und $T_{\max}$ wäre immer kleiner als dieses Zensurereignis. Im Ergebnis würden Verfahren der Ereignisanalyse die Risikomengen unterschätzen und damit Sterberaten überschätzen. Selbst wenn es keine Angaben über Abwanderungen gibt, kann mithilfe eines probabilistischen Modells für die Zwischenereignisse sowie die Abwanderungszeiten über die Verteilung der Lebensdauern nachgedacht werden (Gill 1997; Jonker 2003).

In den Sozialwissenschaften sind Selektionsmodelle eher selten im Zusammenhang mit unvollständigen Angaben in Umfragen oder Registerdaten behandelt worden. Stattdessen dominieren Anwendungen, die sich auf Größen beziehen, die ihre Bedeutung nur im Rahmen eines vorab definierten Modells gewinnen. So untersuchen Diekmann und Wyder (2002) Reputationseffekte bei Internetauktionen, wobei sie auch die erzielten Preise der Auktionen heranziehen. Sie versuchen dabei auch diejenigen Auktionen einzubeziehen, für die gar kein Gebot abgegeben wurde und für die daher auch kein erzielter Preis existiert. Sie argumentieren:

> Die Regressionsschätzung basiert allerdings nur auf der Stichprobe der 99 erfolgreichen Auktionen, da nur für diese ein Verkaufspreis vorliegt. Nun könnte es sich hierbei um ein selektives Sample handeln.... Die Zwei-Stufen-Schätzmethode von Heckman ist eine Alternative, um einen eventuellen Stichprobenauswahlfehler zu kontrollieren (2002: 687).

Ein „Stichprobenauswahlfehler“ könnte aber nur vorliegen, wenn auch den Auktionen ohne Gebote ein „Preis“ zukäme. Diekmann und Wyder unterstellen wohl eine Größe, die mit „Zahlungsbereitschaft“ umschrieben werden kann. Ein Gebot wird abgegeben, falls die Zahlungsbereitschaft eines potentiellen Auktionsteilnehmers größer als das Mindestgebot der Auktion ist. Also fehlen Angaben über die Zahlungsbereitschaft, wenn der Startpreis der Auktion höher als diese Zahlungsbereitschaft ist. Daher könnte das Problem ähnlich wie fehlende Angaben in Umfragen oder Registern behandelt werden. Zwar bleibt unklar, was „Zahlungsbereitschaft“ unabhängig von einem konkreten Gebot in einer gegebenen Auktion bedeuten könnte, sogar, welcher Gruppe von Personen diese Größe zugeschrieben werden soll. Aber selbst wenn dem Konzept eine gewisse Plausibilität zugestanden wird, dann greift der Versuch, ein einfaches Selektionsmodell für fehlende Angaben zu verwenden, zu kurz. Denn zum einen ist bei Auktionen ohne Gebote die Zahlungsbereitschaft kleiner als der Startpreis, so dass die Zahlungsbereitschaft nicht vollständig unbekannt ist. Zum anderen ist der erzielte Preis auch nicht gleich der Zahlungsbereitschaft des Höchstbietenden, sondern (bei mehr als einem Gebot) gleich der Zahlungsbereitschaft des Bieters mit dem zweit höchsten Gebot.[1] Ein Selektionsmodell für „Zahlungsbereitschaft“ müsste beide Aspekte, den der zusätzlichen Information aus den Mindestgeboten und den der erzielten Preise als zweit höchste Zahlungsbereitschaft berücksichtigen. Die Konzentration auf eine Schätzmethode behindert aber oft die Formulierung probabilistischer Modelle für Konzepte, die sich auf fehlende, abgeschnittene oder zensierte und nach Größe selektierte Beobachtungen stützen.

In den Sozialwissenschaften werden Selektionsmodelle auch zur Evaluation sozialpolitischer Maßnahmen und zur Kausalanalyse herangezogen. Die Idee besteht darin, zunächst eine Variable Y festzulegen, die den Erfolg einer Maßnahme oder die Wirkung einer Ursache darstellen soll. Unterstellt wird dann die gleichzeitige Existenz der Variablen (Y_0, Y_1), wobei Y_0 den Wert von Y annimmt, der sich ergeben hätte, wenn jemand nicht an der Maßnahme teilgenommen hätte, Y_1 den bei Teilnahme. Der Erfolg der Maßnahme ließe sich dann etwa durch $Y_1 - Y_0$ ausdrücken.

[1] Genauer: Bietet Person A zunächst maximal 20 Euro und später B 10 Euro, so beträgt der Auktionspreis 10 Euro. Der Auktionspreis gibt also die Zahlungsbereitschaft von B wieder. Bietet andererseits zunächst B 10 Euro, A später 20 Euro, so ist der Auktionspreis 10 Euro plus Mindesterhöhung der Auktion.

Natürlich kann eine Person nur entweder an einer Maßnahme teilnehmen oder nicht teilnehmen. Y_0 und Y_1 können also nicht gleichzeitig beobachtet werden. Ist $\mathcal{Y}$ der Wertebereich der Variablen Y und $\mathcal{Y}^*$ der Wertebereich der beobachteten Variablen Y^*, dann ist

$$\mathcal{Y}^* = (\mathcal{Y} \times \{\mathcal{Y}\}) \cup (\{\mathcal{Y}\} \times \mathcal{Y})$$

Entweder wird der Wert von Y_0 beobachtet, nicht aber der von Y_1, für den nur $y_1 \in \mathcal{Y}$ bekannt ist. Die Beobachtung ist also $(y_0, \mathcal{Y})$. Oder der Wert von Y_1 wird beobachtet, nicht aber der von Y_0. Dann kann die Beobachtung durch $(\mathcal{Y}, y_1)$ angegeben werden. In dieser Formulierung entspricht das Evaluationsproblem einem Problem unvollständiger Angaben. Wird ein passendes probabilistisches Selektionsmodell unterstellt, dann ist die gemeinsame Verteilung von (Y_0, Y_1) identifiziert. Andersherum ist aber klar, dass (Y_0, Y_1) außerhalb dieses Selektionsmodells gar nicht definiert ist. Die Abhängigkeit von willkürlich gesetzten Modellannahmen und die kontrafaktische Formulierung des Evaluationsproblems sind oft kritisiert worden (z.B. Dawid 2000). Die kontrafaktische Formulierung des Problems ist fragwürdig, weil sie einen wohldefinierten Wert für das Ergebnis von Ereignissen voraussetzt, die gar nicht stattgefunden haben. Und im Unterschied zu fehlenden Angaben in Umfragen oder Registerdaten kann die Abhängigkeit der Ergebnisse von den Annahmen des Selektionsmodells nie empirisch ergänzt oder kritisiert werden. Fehlen Angaben in Kirchenregistern, so können Angaben aus Lehns- und Pachtregistern, Handwerksrollen, Sippenbüchern und Gerichtsakten herangezogen werden. Interessiert die Verteilung von Einkommen, dann können Personen, die einmal die Auskunft verweigert haben, nochmals befragt werden. Zudem geben Steuerstatistiken, Sozialversicherungsmeldungen und Lohnstatistiken weitere Auskunft. Aber was in parallelen Welten geschehen würde, in denen alles bis auf die Teilnahme an bestimmten Maßnahmen gleich wäre, entzieht sich jedem Versuch empirischer Überprüfung. Die Betonung der formalen Äquivalenz zwischen beiden Situationen verwischt oft die inhaltlichen Unterschiede. Rubin führt zwei weitere Unterscheidungen an:

> The formal equivalence of these problems … is highly useful coneptually …, but it is, I believe, not helpful to muddle the distinction when trying to generate sound, practical

> statistical advice. The reasons for this conclusion are that (a) the estimands (the things we want to estimate) are fundamentally different for these situations, and (b) the processes that create the missing data are typically very different, both by investigators' design and by nature's devices. (Rubin in Wainer 1992: 183f).

Rubins Punkt (b) verweist zurück auf den Status unvollständiger Daten in den beiden Fällen. Verweigert jemand die Auskunft im Interview, so liegt das in seinem oder ihrem Ermessen, hängt aber nicht von den Interessen und Vorstellungen des Forschers ab. Dagegen sind Daten im kontrafaktischen Modell der Evaluation ebenso wie Überlegungen zur „Zahlungsbereitschaft“ nur unvollständig aufgrund der Modellvorstellungen des Forschers. Erzielte Preise in Auktionen sind ebenso wie Ergebnisse von Arbeitsmarktmaßnahmen zumindest prinzipiell beobachtbar. „Zahlungsbereitschaft“ im Rahmen von Auktionen oder der Erfolg einer Maßnahme (definiert als $Y_1 - Y_0$) ist dagegen nie beobachtbar. Auch wenn diese Unterscheidung eine graduelle ist, so muss doch genau angegeben werden, welche Aussagen getroffen werden sollen. Rubins Punkt (a) soll im Folgenden für unvollständige Angaben in Umfragen diskutiert werden.

4.5. Beispiel: Einkommensangaben im ALLBUS

Das Haushaltsnettoeinkommen ist von zentraler Bedeutung in vielen Bereichen der Sozialforschung, etwa der Armutsforschung und der Haushaltstheorie. Dennoch ist selbst über das mittlere Hausaltsnettoeinkommen empirisch wenig bekannt. Das Statistische Jahrbuch 2001 weist auf der Basis der Einkommens- und Verbrauchsstichprobe (EVS) 1998 einen monatlichen Mittelwert von 5115 DM aus, 5346 DM in Westdeutschland, 4059 DM in Ostdeutschland (Statistisches Bundesamt 2001: 566ff). Dagegen weist der Datenreport 1999 auf der Basis des SOEP für das Jahr 1996 ein mittleres Haushaltsnettoeinkommen von 1978 DM (2061 DM West, 1644 DM Ost) aus (Statistisches Bundesamt 2000: 584). Die beiden Datensätze (EVS und SOEP) unterscheiden sich zwar deutlich:

Die EVS ist eine Quotenstichprobe mit über 60000 beteiligten Haushalten, das SOEP ist eine weit kleinere Zufallsstichprobe.[2] Aber wegen der großen Abweichungen wäre es sicherlich wünschenswert, unabhängigen Aufschluss über die Verteilung des Haushaltseinkommens in der BRD zu erhalten. Von den großen regelmäßigen sozialwissenschaftlichen Umfragen enthält auch der ALLBUS eine Frage nach dem Haushaltsnettoeinkommen sowie nach dem persönlichen Einkommen. Der ALLBUS 1996 wurde als Melderegisterstichprobe durchgeführt. Grundgesamtheit waren Personen ab 18 Jahren in Privathaushalten (einschließlich Deutsch sprechender Ausländer) in West- und Ostdeutschland. Dabei wurden 3518 Interviews realisiert, 2402 davon in den alten und 1116 in den neuen Bundesländern. Personen in den neuen Bundesländern sind also überproportional befragt worden. Ihr Anteil in der Nettostichprobe beträgt 31,7%, der Bevölkerungsanteil betrug ca. 19%. Die berichtete Ausschöpfungsquote betrug 54,2%, d.h. nur 54,2% der angestrebten Interviews wurden realisiert (Koch 2002: 33).[3] Von den 3518 Befragten, die einem Interview zustimmten, antworteten gerade einmal 1772 oder 50,4% auf die Frage nach dem monatlichen Haushaltsnettoeinkommen. Weitere 906 Personen oder 25,8% machten Angaben in gruppierter Form. Insgesamt hat man also nur von etwa 41% der ausgewählten Personen eine valide Antwort auf die Frage nach dem Haushaltseinkommen erhalten, darunter 14% in gruppierter Form. Darüber hinaus gibt es aber noch 293 Personen, die zwar Angaben zu ihrem persönlichen Einkommen machten, aber keine Angaben zum Haushaltseinkommen. Diese Angaben könnten als untere Grenzen für das Haushaltseinkommen benutzt werden. Die Fallzahlen sind in der folgenden Tabelle zusammengestellt.

s_1	Bruttostichprobe	$\lvert s_1\rvert = n_1 = 6491$
s_2	Nettostichprobe	$\lvert s_2\rvert = n_2 = 3518$

[2] Der Materialband zum ersten Armuts- und Reichtumsbericht der Bundesregierung (Bundesregierung 2001) enthält einen hilfreichen Vergleich verschiedener verfügbarer Datenquellen. Fehlende und unvollständige Einkommensangaben im SOEP und ihre Behandlung sind im Datenreport nicht angegeben. Die Arbeiten von Frick und Grapka (2003), Riphahn und Serfling (2002) und Schräpler (2004) geben einen Überblick.

[3] Untersuchungen über den möglicherweise selektiven Ausfall von geplanten Interviews im ALLBUS, dem so genannten Unit-Nonresponse, sind von Koch (1997) und Schneekloth und Leven (2003) vorgelegt worden.

s_3	Haushaltseinkommen angegeben	$\|s_3\| = n_3 = 2678$
s_4	genaue Angaben	$\|s_4\| = n_4 = 1772$
s'_3	Einkommen in [1,9999]	$\|s'_3\| = n'_3 = 2633$
s'_4	genaue Angaben in [1,9999]	$\|s'_4\| = n'_4 = 1748$
s_5	keine Angabe Haushaltseinkommen, aber persönliches Einkommen angegeben	$\|s_5\| = n_5 = 293$
s_6	keine Angabe Haushaltseinkommen, aber genaues persönliches Einkommen	$\|s_6\| = n_6 = 213$

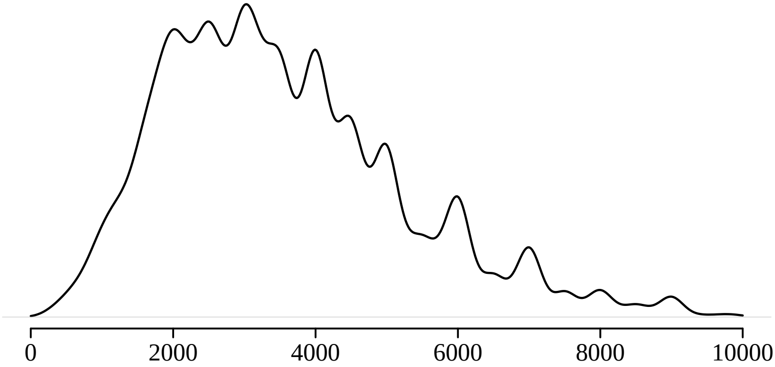


Abbildung 4.1.: Kern-Dichte-Schätzung der genauen Angaben zum Haushaltseinkommen in DM, eingeschränkt auf das Intervall [1,9999]. Als Kern wurde eine Normalverteilung mit Standardabweichung 166 benutzt. Der Dichte-Schätzer ist etwas untergglättet.

Abbildung 4.1 zeigt die Dichte der genauen Angaben zum Haushaltsnettoeinkommen. Diese Angaben sind insbesondere in den höheren Einkommensbereichen zu einem großen Teil auf volle 1000 DM Beträge gerundet. So sind von den 122 Angaben von mehr als 7000 DM 120 in vollen 100 DM Beträgen und 107 in vollen 500 DM Beträgen angegeben, dagegen 85 Angaben in vollen 1000 DM Beträgen. Auf der anderen Seite sind von den 802 Angaben von 3000 DM oder weniger nur 277 in vollen 500 DM Beträgen angegeben. Die Rundungsregeln der Befragten sind unbekannt, hängen aber offenbar von der Einkommenshöhe ab. Auch die „genauen“ Einkommensangaben können nur als grobe Näherung des tatsächlichen verfügbaren Haushaltseinkommen betrachtet werden. Die Rundung in den Einkommensangaben führt zu einer weiteren Unsicherheit bei der Analyse der Daten, die im Folgenden aber nicht systematisch untersucht wird.

Folgt man dem üblichen Verfahren und ignoriert fehlende und gruppierte Angaben, so ergibt sich ein (ungewichtetes) mittleres Haushaltsnettoeinkommen von 3676 DM (3909 DM West, 3171 DM Ost). Beschränkt man sich bei der Berechnung des mittleren Haushaltseinkommens auf Einkommen unter 10000 DM und benutzt nur die genauen Angaben, also die Teilstichprobe s_4', so ergibt sich ein Mittelwert von 3560.[4] Wird das Stichprobendesign ignoriert und eine einfache Zufallsauswahl unterstellt, so kann ein 95%-Konfidenzintervall konstruiert werden: (3480, 3640). Dabei werden die unterschiedlichen Auswahlsätze für Ost- und Westdeutschland unterschlagen, obwohl sich die Mittelwerte in Ostdeutschland (3158 DM) und Westdeutschland (3749 DM) deutlich unterscheiden und ostdeutsche Personen deutlich überrepräsentiert sind. Das geht aber gar nicht anders. Denn weder können die Gewichte für die Teilstichprobe s_2 verwandt werden: dort beträgt der Anteil an Personen in Ostdeutschland 31,7%, während er in der Teilstichprobe s_4' 32,0% beträgt.[5] Noch können Gewichte benutzt werden, die sich aus dem Anteil in der Teilstichprobe s_4' ergeben, denn diese würden sich

[4] Die Beschränkung des betrachteten Einkommenbereichs ist sowohl für nicht-probabilistische wie für probabilistische Modelle notwendig. Für nicht-probabilistische Überlegungen ist das unmittelbar einsichtig, weil Haushalten ohne Einkommensangaben jedes beliebige Einkommen zukommen könnte. Aber auch in probabilistischen Modellen sind Einschränkungen notwendig. Denn ohne solche Einschränkungen ist der Erwartungswert nicht einmal ein stetiges Funktional bezüglich der Kolmogorov-Metrik $d(F, G) \coloneqq \sup_y |F(y) - G(y)|$ zwischen Verteilungsfunktionen F und G (Lehmann 1999: 391). Ohne Einschränkungen existieren also keine gleichmäßig konsistenten Tests, der Bootstrap funktioniert nicht etc. Wird das Verhalten von Schätzern nicht einfach unter einem als ‘wahr’ angenommenen Modell untersucht, sondern werden alle Modelle zugelassen, die mit den gegebenen Beobachtungen statistisch verträglich sind, dann ist der Bereich möglicher Erwartungswerte unendlich groß, selbst wenn die Existenz endlicher vierter Momente unterstellt wird (Davies 1995: 205). Im Zusammenhang mit Selektionsmodellen ist das Problem von Robins und Ritov (1997) untersucht worden (vgl. Abschnitt 4.9).

[5] In der Teilstichprobe s_3' ergibt sich sogar ein Anteil von 34,1%, ein „signifikanter“ Unterschied zu 31,7%.

von Stichprobe zu Stichprobe ändern und erlaubten keine statistischen Aussagen. Man erhält also einen Mittelwert, der wegen der sehr unterschiedlichen Auswahlsätze in Ost- und Westdeutschland wohl nichts über das durchschnittliche Haushaltseinkommen in Deutschland sagt, und ein Konfidenzintervall, dessen Überdeckungseigenschaften schlicht unbekannt sind.[6]

Soll der Rahmen der klassischen Stichprobentheorie nicht vollständig verlassen werden, dann muss die gesamte (realisierte) Stichprobe einschließlich der gruppierten Angaben betrachtet werden. Dies erfordert Verfahren zur Behandlung unvollständiger Angaben. Das einfachste Verfahren hält an der Idee fest, die Angaben der Befragten als fixes Datum zu behandeln. Unvollständige Angaben werden als Bereiche aufgefasst, in denen der tatsächliche Wert liegt. Ist die Auskunft eines Befragten, das Einkommen liege zwischen 2000 DM und 2500 DM, dann wird unterstellt, das exakte Einkommen sei eine der Zahlen 2000, 2001, 2002,…,2499. Bei einer Statistik wie dem Mittelwert wird die Berechnung für jeden der möglichen Werte in diesem Bereich durchgeführt. Das Ergebnis ist ein Bereich von Mittelwerten. Handelt es sich um zusammenhängende Intervalle, dann reicht es, die Statistik für die Extremwerte der Antwortintervalle auszuwerten.[7] Das Verfahren kann nur dann sinnvoll angewandt werden, wenn die Bereiche unvollständiger Daten beschränkt sind. Daher können nur Einkommensangaben etwa unter 10000 DM behandelt werden. Betrachtet man zunächst s_3', so ergibt sich für die möglichen Mittelwerte das Intervall [3589, 3785). Das Intervall der möglichen Mittelwerte ist deutlich länger als das naive

[6] Zur Berechnung eines Gesamtmittelwerts kann auch einfach der bekannte Bevölkerungsanteil in Ost- und Westdeutschland zur Kombination der Ergebnisse in Ost- und Westdeutschland benutzt werden. Dann werden Auswahlsätze einfach ignoriert. Für komplexere Fragen eignet sich eine solche naive Randanpassung allerdings nicht. Zusammenhänge zwischen Stichprobengewichten und Unit-nonresponse werden von Kalton (2002), Kalton und Flores-Cervantes (2003) diskutiert. Little und Vartivarian (2003) kritisieren einige klassische Verfahren.

[7] Die Grundidee ist recht alt (Cochran 1977: Kap. 13.2). Aber schon bei Statistiken wie der Varianz ergeben sich Probleme, effiziente Algorithmen für die Berechnung der Intervalle zu finden (Fishman und Rubin 1998; Rohwer und Pötter 2001: Kap. 19; Ferson et al. 2002). Manski (2003) gibt einen guten Überblick über neuere Ergebnisse. Für Kreuztabellen werden neuere Verfahren von Dobra und Fienberg (2000) beschrieben.

Konfidenzintervall des letzten Absatzes. Um aber die Auswahlwahrscheinlichkeiten des ursprünglichen Stichprobendesigns verwenden zu können, muss zumindest die Teilstichprobe s_2 betrachtet werden, also zusätzlich die 885 Befragten, die gar keine Angaben machten.[8] Dann ergibt sich ein ungewichtetes Intervall der möglichen Mittelwerte von $[2687, 5348)$. Wird gar die Bruttostichprobe s_1 betrachtet, ergibt sich das Intervall $[1457, 7479)$. Für diese Intervalle könnten nun „korrekt" gewichtete Versionen und Konfidenzintervalle ausgerechnet werden. Nur sind die Intervalle selbst schon viel zu groß, um von praktischem Interesse zu sein.

4.6. Stichproben und probabilistische Auswahlmodelle

Da beide Ansätze selten weiterhelfen, wurde versucht, die Fragestellung umzuformulieren. Der hierbei zumeist eingeschlagene Weg opfert einen wesentlichen Ausgangspunkt der Stichprobentheorie, der von den Berichten der Befragten als fixem Datum ausgeht. Stattdessen werden die Angaben der Befragten als Realisationen von Zufallsvariablen im Sinne der Wahrscheinlichkeitstheorie aufgefasst. Für probabilistische Modelle existieren bereits Methoden zur Analyse unvollständiger Daten. Zudem können in diesem Rahmen auch Modelle der Entstehung unvollständiger Angaben entwickelt werden.

Die Durchführung eines solchen Ansatzes ist konzeptionell weit schwieriger und konsequenzenreicher als oft angenommen wird. Einen Teil des Weges gehen *Superpopulationsmodelle*: Sie unterstellen, dass die interessierenden Größen in einer Gesamtheit $\mathcal{U}$ durch einen Zufallsprozess zustande gekommen seien, der sich durch eine Wahrscheinlichkeitsverteilung F_θ beschreiben lässt. Etwa: Das Einkommen der Bevölkerung der BRD wird als Realisation von $N := |\mathcal{U}|$ unabhängigen und identisch log-normalverteilten Zufallsvariablen erzeugt. Das ist offenbar keine realistische Annahme über das Zustandekommen von Einkommen.

[8] Ich rechne die $n_3 - n_3' = 45$ Angaben außerhalb von $[1, 9999]$ dazu.

Die Metaphorik des „als ob durch einen Zufallsprozess zustandegekommen“ erlaubt aber relativ kompakte Beschreibungen von empirischen Verteilungen durch die Parameter θ sowie einen Anschluss an die Stichprobentheorie, denn die realisierten Werte der Zufallsvariablen werden für die Stichprobenziehung als fix angenommen. Der Superpopulationsansatz hält also auf der Ebene der Stichprobenziehung an der Idee der Angaben der Befragten als fixem Datum fest. Allerdings wird das ursprüngliche Problem, Aussagen über die Verteilung eines Merkmals in der Gesamtheit $\mathcal{U}$ zu gewinnen, durch ein anderes ersetzt: Aussagen über θ zu gewinnen. Diese Parameter sind nur durch die Beziehung auf die unterstellte Modellklasse $\{F_\theta \mid \theta \in \Theta\}$ definiert. Ihnen entspricht kein Wert, der sich allein aus der Beobachtung der Werte der Variablen in der Gesamtheit $\mathcal{U}$ gewinnen ließe. Somit wird der realistische Ansatz der klassischen Stichprobentheorie unterlaufen.

Der Superpopulationsansatz geht aber noch nicht weit genug. Wenn man sich für das Antwortverhalten von Befragten interessiert, so müsste im Superpopulationsansatz angenommen werden, dieses Verhalten sei bereits vor jeder Befragung festgelegt, und zwar ganz unabhängig davon, ob jemand tatsächlich befragt wurde oder nicht. Ob also jemand auf die Frage nach dem Haushaltseinkommen gar nicht antwortet oder nur in gruppierter Form, wäre vor jeder Befragung schon entschieden. Denn der Superpopulationsansatz unterstellt fixe Werte (Realisationen des Zufallsprozesses) in der Gesamtheit $\mathcal{U}$ zum Zeitpunkt der Stichprobenziehung. Im nächsten Schritt werden wie in der klassischen Theorie Stichproben aus diesen fixen Werten gezogen. Stichprobenfunktionen wie Mittelwerte hängen auf der Stufe der Stichprobenziehung allein davon ab, wer aus der Gesamtheit $\mathcal{U}$ in die Stichprobe gelangt. Dies ermöglicht den Anschluss an Ergebnisse der klassischen Theorie, hat aber zur Folge, dass das Antwortverhalten aller Mitglieder der Gesamtheit vor jeder Stichprobenziehung festgelegt sein muss.

Soll nicht nur die Tatsache unvollständiger Daten konstatiert, sondern auch ihr Zustandekommen reflektierbar gemacht werden, dann wird es notwendig, auch Variablen wie Interviewform, Merkmale der Interviewer und vieles mehr zu betrachten. Die Annahme, all diese Variablen seien vor jeder Stichprobenziehung für alle Personen der Gesamtheit festgelegt, ist nicht nur fatalistisch und völlig unrealistisch, sondern würde auch die Spezifikation eines Stichprobendesigns wegen der notwendigen Details praktisch unmöglich machen.

Soll an probabilistischen Auswahlmodellen festgehalten werden, dann muss schließlich ganz auf Elemente der Stichprobentheorie verzichtet werden. Sowohl die interessierenden Größen wie das Haushaltseinkommen, die Stichprobenziehung und das Antwortverhalten der Befragten werden in einem einzigen probabilistischen Modell beschrieben. Ist $(\Omega, \mathcal{B}, \lambda)$ ein hinreichend großer Wahrscheinlichkeitsraum, mit dem alle diese Variablen beschrieben werden können, dann lässt sich etwa das Haushaltseinkommen als Funktion von $u \in \mathcal{U}$ und $\omega \in \Omega$ auffassen:

$$Y \colon \mathcal{U} \times \Omega \longrightarrow \{1, 2, 3, \ldots\} =: \mathcal{Y}$$

$Y(u, \omega)$ ist also das Haushaltseinkommen, das einer Person $u \in \mathcal{U}$ bei Realisierung von $\omega \in \Omega$ zukommt. Eine Person u hat in Abhängigkeit von ω verschiedene Einkommen. Aber es gibt ein $\omega_0 \in \Omega$, für das $(Y(u, \omega_0), u \in \mathcal{U})$ den Haushaltseinkommen $(y(u), u \in \mathcal{U})$ in der BRD entspricht.

Der Zusammenhang mit Aussagen über Durchschnitte der $Y(., \omega)$ über alle $u \in \mathcal{U}$ wird hergestellt, indem allen u gleiche Wahrscheinlichkeitsverteilungen zugeschrieben werden und die Unabhängigkeit von $Y(u, .)$ und $Y(u', .)$ für verschieden u, u' angenommen wird.

Tatsächlich in der Stichprobe beobachtet wird aber nur eine mengenwertige Variable mit dem Merkmalsraum

$$\mathcal{Y}^* \coloneqq \{\{y\} \,|\, y \in \mathcal{Y}\} \cup \{[1, 400), [400, 800), \ldots, [15000, \infty)\} \cup \{\mathcal{Y}\}$$

der genaue oder gruppierte Angaben bzw. gar keine Angabe darstellt. Es sei nun

$$S \colon \Omega \longrightarrow \mathcal{P}(\mathcal{U}) \setminus \{\emptyset\}$$

eine Stichprobe, wobei $\mathcal{P}(\mathcal{U})$ die Potenzmenge von $\mathcal{U}$ bezeichnet und S bzgl. $(\Omega, \mathcal{B})$ messbar sein soll. Für die Menge der befragten Personen $u \in S(\omega)$ kann eine neue Variable mit dem Wertebereich $\mathcal{Y}^*$ konstruiert werden, die das „angegebene Haushaltseinkommen“ repräsentiert. Um das Problem zu umgehen, allen Personen unabhängig von der Befragung ein Antwortverhalten zuzuschreiben, wird zu $\mathcal{Y}^*$ noch ein Symbol „*“ hinzugefügt. Dann kann die Definition auf alle Personen $u \in \mathcal{U}$ ausgedehnt werden und es ergibt sich

$$Y^* : \mathcal{U} \times \Omega \longrightarrow \{\{y\} \mid y \in \mathcal{Y}\} \cup \{[1, 400), \ldots, [15000, \infty)\} \cup \{\mathcal{Y}\} \cup \{*\}$$

wobei $Y^*(u, \omega) = *$ für $u \notin S(\omega)$ gesetzt wird. Die Abbildung Y^* repräsentiert das „in der Stichprobe s angegebene Haushaltseinkommen“. Mit dieser Konstruktion ist man nicht gezwungen, über die Antworten nicht befragter Personen zu spekulieren. Sowohl für eine gegebene Stichprobe s, also eingeschränkt auf die Menge $\{\omega \mid S(\omega) = s\}$, als auch auf ganz Ω sind $Y(u, .)$ und $Y^*(u, .)$ Zufallsvariablen im Sinn der Wahrscheinlichkeitstheorie. Der Zusammenhang zwischen $Y(u, .)$ und $Y^*(u, .)$ lässt sich durch probabilistische Modelle darstellen. Sie beziehen sich zunächst auf eine Person u. Es muss zusätzlich angenommen werden, die Zufallsvariablen $(Y(u, .), u \in \mathcal{U})$ bzw. $(Y^*(u, .), u \in \mathcal{U})$ seien stochastisch unabhängig und identisch verteilt. In diesem Rahmen können nun Konsequenzen unvollständiger Angaben für statistische Aussagen abgeschätzt werden.

Die etwas aufwendige Notation ist notwendig, um Verwechslungen zwischen Durchschnitten über die Gesamtheit $\mathcal{U}$ und Verteilungen, Erwartungswerten etc. bezüglich des Wahrscheinlichkeitsraums $(\Omega, \mathcal{B}, \lambda)$ zu vermeiden. In der Literatur erscheint die Verwechslung häufig nach einem nicht kenntlich gemachten Übergang von stichprobentheoretischen zu probabilistischen Argumenten. So schreibt z.B. P. Holland: „A probability will mean nothing more nor less than a proportion of units in $\mathcal{U}$. The expected value of a variable is merely its average value over all of $\mathcal{U}$“ (Holland 1986: 945). Später verwendet er aber die stochastische Unabhängigkeit zwischen Variablen (1986: 948f), ohne zu bemerken, dass stochastisch unabhängige Variablen auf endlichen Räumen $\mathcal{U}$ nur selten existieren. Eine ähnliche Verwechslung findet sich noch bei Vytlacil (2002: 332).

Durchschnitte über $\mathcal{U}$ und Durchschnitte über den Wahrscheinlichkeitsraum $(\Omega, \mathcal{B}, \lambda)$ führen nicht nur zu unterschiedlichen numerischen

Ergebnissen, sie sind nicht einmal konzeptionell verbunden.[9] Zwar garantieren asymptotische Aussagen wie starke Gesetze oder Ergodensätze, dass die beiden Durchschnitte im Grenzwert ($|\mathcal{U}| \rightarrow \infty$ oder $\mathbb{E}(|S|)/|\mathcal{U}| \rightarrow c \notin \{0, 1\}, |\mathcal{U}| \rightarrow \infty$, etc.) gleich sind. Dies sind aber probabilistische Aussagen, die sich auf die Modellebene beziehen, also ein probabilistisches Modell auf $(\Omega, \mathcal{B}, \lambda)$ voraussetzen. Es sind mathematische Konstruktionen, die keine Aussage über empirische Verhältnisse wie Einkommensverteilungen begründen können. Und Grenzwertüberlegungen führen eine zusätzliche Abstraktionsebene ein, die über probabilistische Formulierungen von Antworten auf Fragen nach dem Einkommen hinausgehen. Le Cam und Yang schreiben hierzu:

> It must be pointed out that the asymptotics of the 'standard i.i.d. case' are of little relevance to practical use of statistics, in spite of their widespread study and use. The reason for this is very simple. One hardly ever encounters fixed families $\{p_\theta \mid \theta \in \Theta\}$ with a number of observations that will tend to infinity. There are not that many particles in the visible universe! The use of such considerations is an abuse of confidence that has been foisted upon unsuspecting students and practitioners owing to the fact that we, as a group, possess limited analytical abilities and, perforce, have to limit ourselves to simple problems. …The use of asymptotics 'as $n \rightarrow \infty$' for the standard i.i.d. case seems to be based on an entirely unwarranted act of faith. (Le Cam und Yang 1990: 99f).

Selbst wenn asymptotische Argumente als relevant angesehen werden, so wird man konstatieren müssen, dass unterschiedliche Modelle für F zu

[9] Eine weitere Konsequenz probabilistischer Ansätze betrifft Designvariablen wie die Ost/West-Differenzierung im ALLBUS. Wird eine solche Variable als Konstante bzw. als degenerierte Zufallsvariable aufgefasst, kann sie wie in der klassischen Stichprobentheorie verwandt werden. Insbesondere können Gewichtungsverfahren benutzt werden, um Designaspekte zu berücksichtigen. Werden Designvariablen dagegen als Zufallsvariablen wie alle anderen behandelt, dann hängen diese Variablen von allen anderen Variablen eines Modells ab und Gewichtungsverfahren verlieren ihre Gültigkeit. Die beiden Ansätze führen z.B. bei Regressionsmodellen zu unterschiedlichen Ergebnissen.

Ergebnissen führen können, die offenbar nichts über den Durchschnitt von Werten aller $u \in \mathcal{U}$ sagen.

Die Abwendung von stichprobentheoretischen Konzepten zugunsten probabilistischer Modelle erfordert zudem eine Klärung der Annahmen, die in probabilistischen Modellen verwandt werden. Insbesondere die Annahme unabhängiger und identisch verteilter Zufallsvariablen ist keine Annahme, die verändert oder aufgegeben werden könnte, ohne den Rahmen des Modells zu sprengen. Sie verweist auf keine gesellschaftlichen Sachverhalte, ebenso wenig wie die Leinwand eines Gemäldes auf Eigenschaften der dargestellten Dinge verweist. Entsprechend gibt es auch keine empirischen Anhaltspunkte, aufgrund derer sich die Annahme zurückweisen ließe. Die Annahme ist weder wahr noch falsch, sondern ein Ausgangspunkt für alle probabilistischen Modelle. Wird in einem nächsten Modellierungsschritt eine Modellklasse, z.B. $\{F_\theta \mid \theta \in \Theta\}$ vorgeschlagen, so wird immer schon die Unabhängigkeit und identische Verteilung der so beschriebenen Zufallsvariablen unterstellt. Auch eine Modellklasse kann daher weder wahr noch falsch sein. Aber die Wahl einer Modellklasse kann sich bei einem Vergleich von Realisierungen der Zufallsvariablen mit empirischen Verteilungen als unangemessen erweisen. Eine solche Kritik von Modellvorschlägen, so notwendig und hilfreich sie ist, führt allerdings selbst unter idealisierten Bedingungen nicht zu der eindeutigen Wahl eines probabilistischen Modells. Der spekulative Spielraum, den probabilistische Modelle immer bieten, kann gerade bei der Behandlung von Daten mit unvollständigen Angaben produktiv genutzt werden. Denn dabei muss immer überlegt werden, was der Fall gewesen sein könnte. Probabilistische Modelle bieten einen Rahmen, eine Vielzahl alternativer Möglichkeiten einfach zu benennen und gegeneinander abzuwägen.

4.7. Ignorierbare Ausfälle: Parametrische und nichtparametrische Modelle

Am einfachsten wäre es, wenn eine Angabe $Y^*(u,\omega)$ nur die offensichtliche Information $Y(u,\omega) \in Y^*(u,\omega)$ enthalten würde.[10] Dann bräuchte man sich bei Schätzungen keine Gedanken über den Zusammenhang von $Y(u,.)$ und $Y^*(u,.)$ zu machen. Der Ansatz sei an zwei Beispielen demonstriert: Zunächst sei die Verteilung F_θ von $Y(u,.)$ durch einen endlich-dimensionalen Vektor $\theta \in \mathbb{R}^k$ parametrisiert, etwa $Y(u,.){=}_d\mathrm{N}(\mu,\sigma^2)$, also normalverteilt mit Erwartungswert μ und Varianz σ^2, $\theta = (\mu,\sigma^2) \in \mathbb{R} \times \mathbb{R}_+$. Der Beitrag einer Beobachtung $y(u)$ zur Likelihood $L(\theta; y(u), u \in s)$ ist dann $\phi(y(u);\mu,\sigma^2)$, wobei $\phi(.;\mu,\sigma^2)$ die Dichte der Normalverteilung mit Erwartungswert μ und Varianz σ^2 ist. Sind nun $\{Y(u,.), u \in s\}$ stochastisch unabhängig für gegebenes s, dann ist die Likelihood das Produkt der Dichten

$$L((\mu,\sigma^2); y(u), u \in s) = \prod_{u \in s} \phi(y(u);\mu,\sigma^2)$$

wobei unterstellt wird, der Stichprobenplan habe nichts mit den Variablen $Y(u,.)$ zu tun. Werden alle Informationen über das Zustandekommen einer Realisation $y^*(u)$ von $Y^*(u,.)$ vernachlässigt, wird also nur $y(u) \in y^*(u)$ berücksichtigt, und sind die $Y^*(u,.)$ weiterhin stochastisch unabhängig, dann wird die Likelihood zu[11]

$$L(\theta; y^*(u), u \in s) = \prod_{u \in s} \int_{v \in y^*(u)} dF_\theta(v)$$

[10] Im Folgenden wird unterstellt, dass die Befragten nicht „lügen“, also immer $Y(u,\omega) \in Y^*(u,\omega)$ gilt. Letzteres war, ohne einen probabilistischen Rahmen, bereits bei der Betrachtung intervallwertiger Statistiken unterstellt worden.

[11] Die Formulierung ist bei absolut stetigen Verteilungen wie der Normalverteilung offenbar nicht korrekt. Denn falls es mindestens eine exakte Beobachtung gibt, wird der entsprechende Term 0. Es gibt verschiedene Vorschläge, wie auch absolut stetige Verteilungen in dieser allgemeinen Form behandelt werden können, vgl. Jacobsen und Keiding (1995), Gill et al. (1997) und Nielsen (2000).

Im Fall der Angaben zum Haushaltseinkommen im ALLBUS ergibt sich

$$L(\theta; y^*(u), u \in s_2) = \prod_{u \in s_4} \phi(y(u); \theta) \times \\ \prod_{u \in s_3 \setminus s_4} \int_{v \in y^*(u)} \phi(v; \theta)\, dv \prod_{u \in s_2 \setminus s_3} \int_{v \in \mathbb{R}} \phi(v; \theta)\, dv$$

Der erste Faktor repräsentiert den Beitrag der genauen Beobachtungen, der zweite den der gruppierten Angaben, und der letzte gibt den Beitrag der Verweigerungen (inklusive keine Angabe/weiß nicht) wieder. Der letzte Term ist konstant 1 und damit unabhängig von den Parametern, so dass man sich auf die ersten beiden Terme konzentrieren kann. Maximiert man diese Likelihoodfunktion, ergibt sich für die Nettostichprobe s_2 und ohne Berücksichtigung der unterschiedlichen Auswahlsätze für Ost- und Westdeutschland $\hat{\mu} = 3807$ und $\hat{\sigma} = 2047$ sowie ein modellbasiertes 95%–Konfidenzintervall für μ von (3729, 3885). Der naive Ansatz, der die $n_3 - n_4 = 906$ gruppierten Angaben ganz unberücksichtigt lässt und nur die Daten der Stichprobe s_4 benutzt, ergibt einen Mittelwert von 3676 mit einem (ungewichteten) 95%–Konfidenzintervall (3582, 3770). Der Wert 3676 liegt 131 DM unter dem Wert, der sich unter Berücksichtigung der gruppierten Angaben innerhalb des Normalverteilungsmodells ergibt, sogar ausserhalb des Konfidenzintervalls (3729, 3885).

Anstelle einer Normalverteilung kann auch unterstellt werden, $\log(Y(u, .))$ sei normalverteilt mit den Parametern (μ, σ^2). Die Maximierung der entsprechenden Likelihood führt zu $\hat{\mu} = 8,1113$ und $\hat{\sigma} = 0,5371$. Da $\mathbb{E}_\theta(Y) = \exp(\mu + \sigma^2/2)$ für log-normalverteilte Zufallsvariablen Y ist, ergibt sich als Schätzung des Erwartungswerts 3849 mit dem (modellbasierten) Konfidenzintervall (3765, 3932), berechnet mit der Deltamethode (Lehmann 1999: 85ff). Der Erwartungswert unter diesem Modell ist um 173 DM größer als der naive Durchschnitt. Wählt man schließlich die log-logistische Verteilung mit $F_\theta(y) = \theta y/(1 + \theta y)$, dann ergibt sich $\hat{\theta} = 2,9879 * 10^{-4}$. Der Erwartungswert unter diesem Modell ist allerdings ∞. Man würde auch einen Erwartungswert von ∞ erhalten, wenn die Dichte etwa der log-normalen Verteilung rechts von einem beliebig großen Wert y_0 durch die entsprechende Dichte der log-logistischen Verteilung ersetzt würde. Aber eine solche Veränderung in der Wahl der Modellklasse lässt sich empirisch nicht beurteilen, weil y_0 immer größer als alle Beobachtungen gewählt werden kann. Der Mittelwert von Merkmalen einer endlichen Menge $\mathcal{U}$ ist dagegen sicher immer endlich. Hier zeigt sich der bereits angedeutete Perspektivenwechsel: An Stelle eines Durchschnitts über alle $u \in \mathcal{U}$ interessiert der Durchschnitt über die $\omega \in \Omega$, durch die der Parameter θ erst seine Bedeutung erhält. Es gibt aber zwischen dem Mittelwert einer endlichen Menge $\mathcal{U}$ und dem Erwartungswert einer Zufallsvariablen $Y(u_0, .)$ keine notwendigen Beziehungen.

Es könnte scheinen, das Problem entstehe durch eine zu enge Wahl der Modellklasse und könne durch nichtparametrische Verfahren gelöst werden. Wird „nur“ angenommen, die $(Y(u, .), u \in s)$ seien unabhängig und identisch verteilt, dann ist die nichtparametrische Likelihood für die Verteilung F bei Daten $\{Y^*(u, .) \mid u \in s\}$

$$L(F; y^*(u), u \in s) = \prod_{u \in s} \int_{v \in y^*(u)} dF(v)$$

Im Fall von s_2 ergibt sich

$$L(F; y^*(u), u \in s_2) = \prod_{u \in s_4} F(y(u)) - F(y(u)_-) \times \\ \prod_{u \in s_3 \setminus s_4} \int_{v \in y^*(u)} dF(v) \prod_{u \in s_2 \setminus s_3} \int_{v \in \mathbb{R}} dF(v)$$

wobei $F(y) - F(y_-)$ die Sprunghöhe der Funktion F an der Stelle y ist. Werden nur diskrete Verteilungen F betrachtet, dann existiert häufig ein Maximum der Likelihoodfunktion, etwa $\hat{F}$. $\hat{F}$ wird nichtparametrischer Maximum-Likelihood-Schätzer (NPMLE) der Verteilungsfunktion F genannt. Als Schätzer des Erwartungswerts kann $\hat{\mu} := \int v \, d\hat{F}(v)$ verwandt werden.

Im Fall der Haushaltseinkommen ergibt sich $\hat{\mu} = 3777$. Ein 95%–Konfidenzintervall, basierend auf 10000 Bootstrap-Replikationen, ist (3553, 3859).[12] Auf den ersten Blick könnte es scheinen, als ergäbe sich immer

[12] Die Verteilung der geschätzten Mittelwerte für die Replikationen ist multimodal. Daher ergibt sich ein sehr asymmetrisches Konfidenzintervall, wenn wie hier die Perzentile der Verteilung zur Konstruktion benutzt werden.

ein endlicher Schätzwert $\hat{\mu}$, der nun nicht mehr von der Wahl einer Modellklasse abhinge. Das ist bei unvollständigen Angaben aber nicht der Fall. Der NPMLE ist nicht eindeutig definiert, wenn es gruppierte Beobachtungen gibt, aber keine genauen Beobachtungen in dieses Intervall fallen. Dann kann $\hat{F}$ auf dem Intervall beliebig definiert werden, ohne den Wert der Likelihoodfunktion zu ändern. Fällt insbesondere keine genaue Beobachtung in die größte Einkommensklasse „Einkommen größer als 15000 DM" während das Intervall wenigstens einmal genannt wird, dann kann die zugehörige Masse in $\hat{F}$ beliebig gegen ∞ verschoben werden. Das geschieht in den Bootstrap-Replikationen so oft, dass die obere Schranke des Konfidenzintervalls ehrlicherweise durch ∞ ersetzt werden müsste. In der Tat ist in der Berechnung des „Konfidenzintervalls" der replizierte Datensatz einfach die Angabe „Einkommen größer als 15000 DM" durch den Wert 22500 ersetzt worden, falls keine genauen Beobachtungen in das Intervall fielen. Jeder andere Wert > 15000 wäre aber genauso möglich. Auch unter nichtparametrischen Modellen ergibt sich ein beliebig großer Unterschied zwischen dem ursprünglich interessierenden Durchschnitt von Werten einer Gesamtheit $\mathcal{U}$ und dem Erwartungswert $\int v\,dF(v) = \int Y(u_0,\omega)\,d\lambda(\omega)$, einem Durchschnitt über $\omega \in \Omega$.

4.8. Ignorierbare Ausfälle: MAR, CAR und all das

Der Perspektivenwechsel führt zu einer Abkehr von dem Versuch, probabilistische Aussagen über den Zusammenhang eines Mittelwerts in einer Gesamtheit $\mathcal{U}$ mit Mittelwerten über Stichproben zu formulieren. Das mag gerechtfertigt sein, wenn stattdessen die Verwendung probabilistischer Modelle zum Verständnis von Effekten unvollständiger Angaben beiträgt, zumal über deren Zustandekommen nur spekuliert werden kann. Zunächst muss geklärt werden, unter welchen Bedingungen die Verwendung von $L(\theta; y^*(u), u \in s)$ im vorigen Abschnitt begründet werden kann. Dies kann nicht immer der Fall sein, selbst wenn für alle Personen $u \in \mathcal{U}$ identische Beziehungen zwischen $Y(u,.)$ und $Y^*(u,.)$ unterstellt werden.

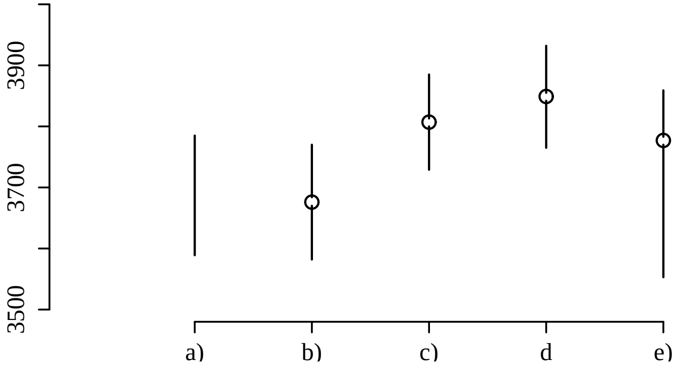


Abbildung 4.2.: Geschätztes mittleres Haushaltseinkommen und 95% Konfidenzintervalle. a) Nichtparametrische Schranken auf s_3', b) Normalverteilung ohne gruppierte Angaben, c) Normalverteilung unter Einschluss gruppierter Angaben, d) Lognormalverteilung unter Einschluss gruppierter Angaben, e) Nichtparametrischer Mittelwert unter Einschluss gruppierter Angaben

Sei z.B. $\mathcal{Y} = \{1000, 1500, 2000\}$ und $\mathcal{Y}^* = \{\{1000, 1500\}, \{1000\}, \{1500\}, \{2000\}\}$. Ist nun $y(u) = 1000$, dann kann u entweder $\{1000\}$ berichten, also den genauen Betrag nennen, oder aber mit $\{1000, 1500\}$ antworten. Entsprechendes gilt für $y(u) = 1500$. Ist nun $Y(u, .)$ auf $\{1000, 1500, 2000\}$ gleichverteilt und berichten alle Personen $y^*(u) = \{1000, 1500\}$ falls $y(u) = 1000$, sonst aber immer den genauen Betrag, dann ergibt sich die folgende Verteilung auf $\mathcal{Y}^*$:

$$\begin{aligned}
&\Pr(Y^*(u, .) = \{1000, 1500\}) = 1/3\\
&\Pr(Y^*(u, .) = \{1000\}) = 0\\
&\Pr(Y^*(u, .) = \{1500\}) = 1/3\\
&\Pr(Y^*(u, .) = \{2000\}) = 1/3
\end{aligned}$$

Werden die Beobachtungen wie im letzten Abschnitt behandelt, so wird

$$\Pr(Y^*(u, .) = \{1000, 1500\}) = \Pr(Y(u, .) = 1000) + \Pr(Y(u, .) = 1500)$$

gesetzt. Ist die Verteilung von $Y^*(u, .)$ bekannt, so ergibt sich als Verteilung von $Y(u, .)$:

$$\Pr(Y(u, .) = 1000) = 0, \ \Pr(Y(u, .) = 1500) = 2/3,$$

$$\Pr(Y(u,.) = 2000) = 1/3$$

Für den Zusammenhang zwischen Y und Y^* wird

$$\begin{aligned}&\Pr(Y^*(u,.) = \{1000, 1500\} \mid Y(u,.) = 1000) = 1/2\\ &= \Pr(Y^*(u,.) = \{1000, 1500\} \mid Y(u,.) = 1500)\end{aligned}$$

angenommen. Die bedingten Wahrscheinlichkeiten von $\{Y^*(u,.) = \{1000, 1500\}\}$ sind für beide möglichen Bedingungen $\{Y(u,.) = 1000\}$ und $\{Y(u,.) = 1500\}$ gleich. Kombiniert man die unterstellte Verteilung von $Y(u,.)$ mit den beiden konditionalen Verteilungen, dann ergibt sich in der Tat die Verteilung von $Y^*(u,.)$. Die Interpretation wäre: Ist $Y(u,.) = 1500$, dann antwortet u mit Wahrscheinlichkeit 1/2 entweder $\{1500\}$ oder $\{1000, 1500\}$. Die entscheidende Annahme ist offenbar die über die bedingten Verteilungen

$$\begin{aligned}&\Pr(Y^*(u,.) = \{1000, 1500\} \mid Y(u,.) = 1000) \text{ und}\\ &\Pr(Y^*(u,.) = \{1000, 1500\} \mid Y(u,.) = 1500)\end{aligned}$$

Die bedingten Verteilungen beschreiben den Zusammenhang zwischen $Y(u,.)$ und $Y^*(u,.)$.

Sei nun allgemein $Y(u,.)$ eine Zufallsvariable mit Werten in der endlichen Menge $\mathcal{Y}$ und $Y^*(u,.)$ eine Zufallsvariable mit Werten in $\mathcal{Y}^* \subseteq \mathcal{P}(\mathcal{Y}) \setminus \{\emptyset\}$ auf dem gemeinsamen Raum $(\Omega, \mathcal{B}, \lambda)$. Dann soll $Y^*(u,.)$ *zufällige Vergröberung* (CAR, coarsened at random) von $Y(u,.)$ heißen, wenn eine der folgenden äquivalenten Bedingungen für alle $y^* \in \mathcal{Y}^*$ und für alle $y \in y^*$ erfüllt ist:

$$\Pr(Y^*(u,.) = y^* \mid Y(u,.) = y) \text{ ist konstant auf } y \in y^* \tag{4.1}$$

$$\Pr(Y^*(u,.) = y^* \mid Y(u,.) = y) = \Pr(Y^*(u,.) = y^* \mid Y(u,.) \in y^*) \tag{4.2}$$

$$\{Y^*(u,.) = y^*\} \perp\!\!\!\perp \{Y(u,.) = y\} \mid \{Y(u,.) \in y^*\} \tag{4.3}$$

$$\Pr(Y(u,.) = y \mid Y^*(u,.) = y^*) = \Pr(Y(u,.) = y \mid Y(u,.) \in y^*) \tag{4.4}$$

Dabei bedeutet $A \perp\!\!\!\perp B \mid C$ die bedingte stochastische Unabhängigkeit der Ereignisse A und B gegeben C. $\{\Pr(Y^*(u,.) = y^* \mid Y(u,.) = y) \mid y* \in \mathcal{Y}^*, y \in y^*\}$ soll im Folgenden *Selektionsmodell* heißen.

Die Bedingungen (4.1) und (4.2) beschreiben die Situation ausgehend vom Wert $y(u)$ der zugrunde liegenden Variablen $Y(u, .)$. Bei gegebenem $y(u)$ müssen nur noch die Antwortmöglichkeiten $Y^*(u, .)$ in Erwägung gezogen werden. Ist etwa das tatsächliche Haushaltseinkommen eines Befragten 1234 DM, dann kann er im ALLBUS entweder 1234 oder das Intervall [1000, 1250) angeben, oder er antwortet gar nicht. Aus diesen Antwortalternativen wählt u mit den bedingten Wahrscheinlichkeiten $\Pr(Y^*(u, .) = y^* \mid Y(u, .) = 1234)$. Ist dagegen $y(u) = 1123$, so erzwingt (4.1) eine Auswahl zwischen den drei Antwortmöglichkeiten 1123, [1000,1250) und $[1, \ldots, \infty)$ mit den gleichen Wahrscheinlichkeiten wie im Fall $y(u) = 1234$. Für alle möglichen Einkommen im Bereich [1000, 1250) entscheidet sich u nach den gleichen Wahrscheinlichkeiten zwischen einer genauen Angabe, der gruppierten Angabe oder gar keiner Angabe. Für Werte in einem anderen Gruppierungsintervall kann sich eine andere Aufteilung zwischen den Antwortmöglichkeiten „genau“ und „gruppiert“ ergeben. Die Bedingung (4.2) ist nur eine Umformulierung dieser Beschreibung. Denn (4.2) verlangt, dass die Entscheidung, kein Haushaltseinkommen anzugeben, unabhängig von dem tatsächlichen Einkommen getroffen wird, während die Entscheidung zwischen einer gruppierten oder genauen Angabe innerhalb eines Gruppierungsintervalls nicht von der tatsächlichen Höhe des Einkommens abhängt. Wenn bekannt ist, dass das tatsächliche Einkommen $Y(u, .) \in [1000, 1250)$ ist, dann verlangt (4.3) die Unabhängigkeit des Ereignisses „gruppierte Angabe“ von den tatsächlichen Einkommen innerhalb des Intervalls. Ist $y^* = \mathcal{Y} = [1, 2, \ldots, \infty)$, dann ist die Bedingung $Y(u, .) \in [1, 2, \ldots, \infty)$ immer erfüllt und (4.3) verlangt die (unbedingte) stochastische Unabhängigkeit von $\{Y^*(u, .) = [1, 2, \ldots, \infty)\}$ und $Y(u, .)$. Ist dagegen $Y(u, .) \in y^* = \{y\}$, dann gilt $Y(u, .)) = y$ und die Bedingung (4.3) ist automatisch erfüllt.

Die Bedingung (4.4) ist besonders hilfreich, weil sie nicht von den zugrunde liegenden Werten $Y(u, .)$ sondern von den beobachteten Angaben $Y^*(u, .)$ ausgeht. Ist etwa $y^*(u) = [1000, 1250)$, dann verlangt (4.4), dass die Verteilung der tatsächlichen Werte $Y(u, .)$ sich nicht von der bedingten Verteilung der $Y(u, .)$ unterscheidet, wenn bekannt ist, dass $Y(u, .)$ in dem Intervall [1000,1250) liegt. Genau dies ist in der Konstruktion der Likelihoodfunktionen im letzten Abschnitt verwandt worden.

Schreibt man $\Pr_\theta(Y(u,.) = y)$, um die Abhängigkeit der Verteilung von einem Parameter $\theta \in \Theta$ anzugeben, und entsprechend $\Pr_\gamma(Y^*(u,.) = y^* \mid Y(u,.) = y)$ mit $\gamma \in \Gamma$ für das Selektionsmodell, dann kann die Verteilung der Beobachtungen $Y^*(u,.)$ wie folgt aufgespalten werden:

$$\begin{aligned}
&\Pr{}_{\theta,\gamma}(Y^*(u,.) = y^*) \\
&= \sum_{y \in y^*} \Pr{}_{\theta,\gamma}(Y^*(u,.) = y^*, Y(u,.) = y) \\
&= \sum_{y \in y^*} \Pr{}_{\theta}(Y(u,.) = y) \Pr{}_{\gamma}(Y^*(u,.) = y^* \mid Y(u,.) = y) \\
&= \Pr{}_{\gamma}(Y^*(u,.) = y^* \mid Y(u,.) = y, y \in y^*) \sum_{y \in y^*} \Pr{}_{\theta}(Y(u,.) = y) \\
&= \Pr{}_{\theta}(Y(u,.) \in y^*) \Pr{}_{\gamma}(Y^*(u,.) = y^* \mid Y(u,.) \in y^*)
\end{aligned}$$

Die dritte Gleichung folgt aus der CAR-Bedingung (4.1), die letzte Gleichung aus (4.2). Sind die Parameter θ und γ variationsunabhängig, gibt es also zu jedem Element $(\theta, \gamma) \in \Theta \times \Gamma$ eine Wahrscheinlichkeitsverteilung $\Pr_{\theta,\gamma}$, dann kann bei Likelihoodbetrachtungen für θ der Selektionsteil $\Pr_\gamma(Y^*(u,.) = y^* \mid Y(u,.) \in y^*)$ vernachlässigt werden. Es reicht,

$$L(\theta; y^*(u), u \in s) = \prod_{u \in s} \Pr{}_{\theta}(Y(u,.) \in y^*)$$

zu maximieren. Antworten die Befragten entweder mit einer genauen Angabe oder gar nicht, dann ist $\mathcal{Y}^* = \{\{y\} \mid y \in \mathcal{Y}\} \cup \{\mathcal{Y}\}$. In diesem Fall impliziert CAR die Möglichkeit, sich nur auf die genauen Angaben beschränken zu können. Die CAR-Bedingung ist in diesem Zusammenhang auch MAR (missing at random) genannt worden.

Es kann gezeigt werden, dass es zu jeder vorgelegten Verteilung auf $\mathcal{Y}^*$ immer eine Verteilung auf $\mathcal{Y}$ und ein Selektionsmodell gibt, der die CAR-Bedingung erfüllt (Gill et al. 1997: 262; Heitjan 1994; Heitjan und Rubin 1991; Grünwald und Halpern 2003). Ist insbesondere $\{y\} \in \mathcal{Y}^*$ für alle $y \in \mathcal{Y}$ und $\Pr(Y^*(u,.) = \{y\}) > 0$, dann ist die Verteilung von $Y(u,.)$ durch die CAR-Bedingung sogar eindeutig bestimmt. Ist die

Verteilung von $Y^*(u,.)$ bekannt, dann kann immer ein CAR-Modell unterstellt werden. Mit anderen Worten: Keine noch so große Menge an Daten erlaubt es, zwischen einem ignorierbaren Selektionsmodell, der die CAR-Bedingung erfüllt, und nicht ignorierbaren Modellen (wie am Anfang des Abschnitts) zu unterscheiden.[13] Selektionsmodelle sind nicht identifizierbar: Wird ein beliebiges Selektionsmodell vorgeschlagen, so kann immer ein CAR-Modell angegeben werden, der ebenso gut zu den Daten passt.

Wenn von einem CAR-Modell ausgegangen wird, dann braucht, basierend auf der Likelihoodtheorie, kein Selektionsmodell angegeben zu werden. Manchmal erscheint es aber sinnvoll, sich selbst in der CAR-Situation ein Bild des Selektionsprozesses zu machen. Wird ein (semi-) parametrisches Modell für den Selektionsprozess gewählt, dann hat dies empirische Konsequenzen, kann sich also als falsch erweisen. Denn unter der CAR-Bedingung sind auch die $\Pr(Y^*(u,.) = y^* \mid Y(u,.) \in y^*)$ eindeutig bestimmt, falls nur $\Pr(Y^*(u,.) = y^*) > 0$ ist. Die Annahme einer Klasse $\Pr_{\gamma}(Y^*(u,.) = y^* \mid Y(u,.) \in y^*)$ von Selektionsmodellen kann unter der CAR-Bedingung zumindest potentiell aus empirischen Gründen zurückgewiesen werden, jedenfalls dann, wenn neben exakten und vollständig fehlenden Angaben auch partielle Angaben zur Verfügung stehen und modelliert werden.

Zusammenfassend lässt sich sagen, dass erstens probabilistische Überlegungen zu einer nicht trivialen Charakterisierung von Bedingungen führen, unter denen klassische Methoden für unvollständige Beobachtungen korrekt sind: die CAR-Bedingungen. Zweitens zeigt sich, dass Selektionsmodelle empirisch nicht identifiziert sind: Es kann immer ein CAR-Modell konstruiert werden, das die Daten exakt reproduziert, ganz unabhängig davon, wie die Daten „tatsächlich“ entstanden sind. Die CAR-Annahme und damit die Verwendung klassischer Likelihood-Methoden lässt sich empirisch nicht hinterfragen. Drittens ist es selbst unter der CAR-Bedingung möglich, einige (semi-) parametrische Selektionsmodelle empirisch zurückzuweisen, wenn neben exakten und

[13] Das Ergebnis gilt nicht nur für endliche Mengen $\mathcal{Y}$, sondern im wesentlichen auch in allgemeinen Räumen. Allerdings wird dann die Formulierung sehr aufwendig (Gill et al. 1997: 273ff).

vollständig fehlenden Angaben auch gruppierte oder andere partielle Angaben vorliegen und entsprechend modelliert werden. Man darf aber bei all dem Fortschritt nicht vergessen, dass probabilistische Modelle untersucht werden. Die Modelle können nicht umstandslos mit dem realen Verhalten von Befragten gleichgesetzt werden. Denn die Modelle unterstellen u.a., alle Befragten würden ihr Antwortverhalten nach einer Wahrscheinlichkeitsverteilung auswürfeln, die zudem für alle gleich wäre.

4.9. Ignorierbare Ausfälle: Konditionale CAR-Modelle

In vielen Fällen gibt es neben den Angaben $(y^*(u), u \in s)$ zu den interessierenden Größen $(y(u), u \in \mathcal{U})$ weitere Informationen. Dabei kann es sich um Designvariablen handeln, deren Werte zumindest für die intendierte Stichprobe bekannt sind, um Angaben über die Kontaktaufnahme oder um Angaben des Befragten aus anderen Teilen des Interviews. Im ALLBUS gibt es z.B. Angaben für alle Befragten aus s_2 zu Geschlecht, Alter, Haushaltsgröße, Staatsangehörigkeit und Befragungsgebiet (Ost/West). Nun kann die globale Annahme einer CAR-Bedingung unrealistisch erscheinen. Gleichwohl könnte die CAR-Bedingung getrennt für alle Teilmengen gelten, die durch die zusätzlichen Angaben $X(u, .)$ definiert werden.

Wenn die zusätzlichen Angaben in einem Vektor X zusammengefasst werden, dann kann ein entsprechender Zufallsvektor konstruiert werden: $X : \mathcal{U} \times \Omega \to \mathcal{X}$. Die CAR-Bedingungen können konditional auf die Werte dieser Kovariablen formuliert werden:

$$\Pr(Y^*(u, .) = y^* \mid Y(u, .) = y, X(u, .) = x) \text{ ist konstant auf } y \in y^* \quad (4.5)$$

$$\begin{aligned} &\Pr(Y^*(u, .) = y^* \mid Y(u, .) = y, X(u, .) = x) \\ &\qquad = \Pr(Y^*(u, .) = y^* \mid Y(u, .) \in y^*, X(u, .) = x) \quad (4.6) \end{aligned}$$

$$\{Y^*(u,.) = y^*\} \;\perp\!\!\!\perp\; \{Y(u,.) = y\} \;\mid\; \{Y(u,.) \in y^*\}, X(u,.) \quad (4.7)$$

$$\begin{aligned}\Pr(Y(u,.) = y \mid Y^*(u,.) = y^*, X(u,.) = x)&\\ = \Pr(Y(u,.) = y \mid Y(u,.) \in y^*, X(u,.) = x)& \quad (4.8)\end{aligned}$$

In vielen Texten wird suggeriert, die CAR-Annahme werde plausibler, wenn nur genügend „Informationen" in Form von Kovariablen in das Selektionsmodell einbezogen werden, wenn also ein möglichst großer Vektor $X(u,.)$ gewählt wird. So schreiben Little und Rubin:

> ...we believe that in situations where good covariate information is available and included in the analysis, the missing at random (MAR) assumption may often be a reasonable approximation to reality, thus obviating the need for a sensitivity analysis to model nonignorable nonresponse. (Little und Rubin in Scharfstein et al. 1999: 1130).

Die Argumentation beruht auf einer fehlerhaften Gleichsetzung von „Information" mit bedingten Verteilungen. Während im umgangssprachlichen Gebrauch des Wortes eine „bessere Information" immer zu einem besseren Verständnis einer Situation oder eines Ereignisses beiträgt, gilt dies nicht für die Einbeziehung zusätzlicher Kovariabler in den Bedingungen (4.5) – (4.8).[14]

Eine Idee, die schon von Pearson Anfang des letzten Jahrhunderts formuliert wurde, geht davon aus, dass Y eine lineare Regression auf den

[14] Sind z.B. U, V normalverteilte unabhängige Zufallsvariablen mit Erwartungswert 0 und Varianz 1, dann sind $Y_1 := U+V$ und $Y_2 := U-V$ unabhängig und normalverteilt, also $Y_1 \perp\!\!\!\perp Y_2$. Dagegen ist die bedingte Kovarianz, wenn die „Information" $U = u$ gegeben ist: $\mathrm{Cov}(Y_1, Y_2 \mid U = u) = \mathbb{E}(Y_1 Y_2 \mid U = u) - \mathbb{E}(Y_1 \mid U = u)\mathbb{E}(Y_2 \mid U = u) = \mathbb{E}((u+V)(u-V)) - u^2 = \mathrm{Var}(V) - u^2 = 1 - u^2$. $Y_1 \perp\!\!\!\perp Y_2 | U = u$ gilt also nur, falls $U = 0$ ist, ein Ereignis mit Wahrscheinlichkeit 0. Die Einführung der zusätzlichen „Information" $U = u$ führt von unabhängigen Variablen Y_1 und Y_2 zu korrelierten bedingten Variablen. Die Bedingung (4.7) kann durch die Einbeziehung weiterer Kovariabler also auch verletzt werden. Weitere Probleme bei der Interpretation von „Information" bei bedingten Modellen diskutieren Dubra und Echenique (2004).

Vektor X besitzt, $\mathbb{E}(Y \mid X) = X\beta$ (Lawley 1943). Die lineare Beziehung soll dabei sowohl für die Teilmenge der vollständigen Angaben als auch für die Gesamtheit $\mathcal{U}$ mit dem jeweils gleichen β gelten. An Stelle der Identität der Erwartungswerte in allen Teilstichproben wird also nur die Identität der Regressionsfunktion sowie eine konstante bedingte Varianz gefordert. Haben $Y \mid X$ und $Y \mid X, \mathbb{1}[R = 1]$ die gleiche Verteilung, dann ist das Modell sicherlich CAR. Wird nur die Gleichheit der Erwartungswerte und Varianzen gefordert, so ergibt sich ein etwas allgemeineres Selektionsmodell.

Im Pearson-Lawley Modell kann zunächst β auf der Teilstichprobe mit vollständigen Angaben geschätzt werden; in einem zweiten Schritt aufgrund von $\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y \mid X))$ ein Schätzer des Erwartungswerts von Y durch

$$\hat{\mu} = \int \hat{\mathbb{E}}(Y \mid X = x)\, d\hat{F}_X(x) = \frac{1}{|s|} \sum_{u \in s} x(u)\hat{\beta}$$

konstruiert werden. Ist die Regression auch homoskedastisch, dann lassen sich Schätzer für die Varianzen angeben. Im Fall des ALLBUS ergibt sich $\hat{\mu} = 3763$ mit dem 95% Konfidenzintervall (3660, 3866), wenn als Kovariablen das Alter, Geschlecht, Haushaltsgröße und Staatsangehörigkeit benutzt werden.[15] Der naive Mittelwert, der nur vollständige Angaben berücksichtigt, ist um 90 DM kleiner und befindet sich am unteren Rand des Konfidenzintervalls.

Im Rahmen des Pearson-Lawley Ansatzes können auch die gruppierten Angaben aus $s_3 \setminus s_4$ berücksichtigt werden, indem die Methoden des letzten Abschnitts auf die Residuen $y(u) - x(u)\beta$ angewandt werden. Außerdem kann an Stelle der linearen Regression ein beliebiges (parametrisches oder semiparametrisches) Regressionsmodell benutzt werden. Ein Nachteil der Methode ist aber ihre Abhängigkeit von Annahmen

[15] Die 6 Beobachtungen ohne Altersangabe wurden ausgeschlossen, die Schätzung von β erfolgte auf s_4 ohne diese Beobachtungen. Die Haushaltsgröße bezieht sich auf die Anzahl der Personen im Haushalt, einschließlich der Kinder. Die Angabe wurde gruppiert, indem Haushalten mit mehr als 4 Personen der Wert 5 zugeordnet wurde. Bei der Staatsangehörigkeit wird nur unterschieden, ob jemand einen ausländischen Pass hat oder nicht.

bezogen auf die Regressionsgleichung. Diese ist in der Regel nicht direkt von Interesse. Im Fall des ALLBUS soll eine Aussage über das mittlere Haushaltseinkommen getroffen werden, nicht aber über einen Regressionszusammenhang. Zudem sind die Kovariablen X, die für alle Befragten zur Verfügung stehen, hauptsächlich durch das Stichprobendesign und die Fragebogenkonstruktion bestimmt, nicht durch inhaltliche Überlegungen. Daher wird versucht, den Regressionszusammenhang $\mathbb{E}(Y \mid X)$ möglichst allgemein, d.h. ohne parametrische Annahmen, zu modellieren. Das Haushaltseinkommen ist sicherlich keine lineare Funktion des Alters. Theoretisch gibt es keinen sinnvollen Zusammenhang zwischen dem Alter eines Befragten und dem Haushaltseinkommen, es sei denn, es handelt sich um Ein-Personen-Haushalte. Der lineare Term für das Alter sollte daher flexibler, etwa durch Spline-Funktionen dargestellt werden. Zudem sollten möglichst alle Interaktionen zwischen den Kovariablen berücksichtigt werden. Wenn aber keine parametrischen Annahmen oder wenigstens Annahmen über die Glätte der Regressionsbeziehung getroffen werden können, dann gibt es innerhalb des Ansatzes nicht einmal ein Schätzverfahren, das gleichmäßig konsistent ist (Robins und Ritov 1997: 294f). Selbst wenn diese theoretischen Schwierigkeiten ignoriert werden, bleibt das praktische Problem, ein relativ stabiles und gleichzeitig allgemein akzeptierbares Regressionsmodell für Y gegeben X zu formulieren.

Eine Möglichkeit, zumindest einige der theoretischen Probleme zu umgehen, ergibt sich aus einem Rückgriff auf eine Idee der Stichprobentheorie und verwendet gewichtete Schätzgleichungen. Wird nur zwischen vollständigen und fehlenden Angaben unterschieden, dann kann ein probabilistisches Modell für das Fehlen einer Angabe in Abhängigkeit von den Kovariablen X formuliert werden. Wird $R(u,.) = 1$ gesetzt, falls $Y(u,.)$ beobachtet wurde, $R(u,.) = 2$ sonst, dann ist ein Modell für das Fehlen von Angaben über $Y(u,.)$ etwa durch $\pi(u,x) \coloneqq \Pr(R(u,.) = 1 \mid X(u,.) = x) = \Pr(R(u,.) = 1 \mid X(u,.) = x, Y(u,.) = y)$ bestimmt. Die Variable $\pi(u,X)$ wird häufig „Propensity Score“ genannt. Ignorierbarkeit des Selektionsmodells besteht gerade in der Unabhängigkeit des Propensity Scores $\pi(u,x)$ von y, also in der Annahme $R(u,.) \perp\!\!\!\perp Y(u,.) \mid X(u,.)$. Ist die Auswahlwahrscheinlichkeit für alle Kovariablenwerte größer als eine positive Schranke, $\pi(u, X) > \sigma > 0$ für alle $X(u, .)$, dann gilt

$$\begin{aligned}
&\mathbb{E}\left(\frac{\mathbb{1}[R(u,.)=1]Y(u,.)}{\pi(u,X)}\right)\\
&\quad=\mathbb{E}\left(\mathbb{E}\left(\frac{\mathbb{1}[R(u,.)=1]Y(u,.)}{\pi(u,X)}\,\middle|\, Y(u,.)=y, X(u,.)=x\right)\right)\\
&\quad=\mathbb{E}\left(\frac{Y(u,.)}{\pi(u,X)}\mathbb{E}\left(\mathbb{1}[R(u,.)=1] \mid Y(u,.)=y, X(u,.)=x\right)\right)\\
&\quad=\mathbb{E}\left(\frac{\pi(u,X)Y(u,.)}{\pi(u,X)}\right)=\mathbb{E}\left(Y(u,.)\right)
\end{aligned}$$

Daher ist

$$\hat{\mu}=\sum_{u\in s}\frac{\mathbb{1}[R(u,.)=1]Y(u,.)}{\pi(u,X)}\Big/\sum_{u\in s}\frac{\mathbb{1}[R(u,.)=1]}{\pi(u,X)}$$

ein erwartungstreuer Schätzer des Erwartungswerts der $Y(u, .)$, und zwar ganz unabhängig von der Form des bedingten Erwartungswerts $\mathbb{E}(Y \mid X)$. Allerdings muss ein Schätzer für $\pi(u, X)$ angegeben werden. Wird z.B. ein Logit-Modell $\hat{\pi}(u, x(u)) = \exp(x(u)\hat{\beta})/1 + \exp(x(u)\hat{\beta})$ verwendet, dann ergibt sich für das mittlere Haushaltseinkommen $\hat{\mu} = 3765$ mit einem Konfidenzintervall von (3599, 3931).

Das deutlich größere Konfidenzintervall des gewichteten Schätzers im Vergleich zum Pearson-Lawley-Schätzer ist eine Konsequenz sowohl der abgeschwächten probabilistischen Annahmen als auch der nur unvollständigen Nutzung der Verteilung der $X(u, .)$ sowie der gemeinsamen Verteilung von $Y(u, .)$ und $X(u, .)$ auf $\{u \in s \mid R(u, .) = 1\}$ für die Schätzung. Ein Teil der Information kann durch die Addition eines weiteren, von Funktionen der $X(u, .)$ und $\mathbb{1}[R(u, .) = 1]Y(u, .)$ abhängigen Terms zur Schätzgleichung zurückerhalten werden.[16] Allerdings können im

[16] Verschiedene Varianten der Einbeziehung dieser Information sind sowohl von Qin et al. (2002) als auch von Robins und seinen Mitarbeitern untersucht worden. Theoretisch können „optimale“ erweiterte Schätzfunktionen angegeben werden, die die (asymptotische) Varianz von $\hat{\mu}$ minimieren. Die optimale Wahl einer Schätzfunktion hängt von der Spezifikation eines Modells der bedingten Verteilung von $Y(u, .)$ gegeben $X(u, .)$ bzw. gegeben $X(u, .), R(u, .) = 1$ ab (Rotnizky und Robins 1997;

Rahmen gewichteter Schätzer partielle Angaben nur sehr rudimentär Berücksichtigung finden. Die Schätzgleichung beruht zunächst nur auf vollständigen Angaben, partielle Angaben werden im Gegensatz zum Pearson-Lawley Ansatz oder zu Likelihood-Methoden nur über weitere additive Terme in die Schätzgleichung eingeführt.

4.10. Nicht ignorierbare Ausfälle

Bei der Diskussion von Selektionsmodellen wird überlegt werden müssen, was geschieht, wenn sie nicht die CAR-Bedingung erfüllen. In diesem Fall hängt $\Pr(Y^*(u,.) = y^* \mid Y(u,.) = y)$ aufgrund von (4.1) von $y \in y^*$ ab und diese bedingten Wahrscheinlichkeiten könnten modelliert werden. Andersherum kann von (4.4) ausgegangen und entsprechend $\Pr(Y(u,.) = y \,|\, Y^*(u,.) = y^*)$ modelliert werden. Eine Reihe anderer Möglichkeiten der Konstruktion nicht ignorierbarer Selektionsmodelle sind denkbar. Das wohl bekannteste Modell ist von Heckman (1976) vorgeschlagen und zumeist für den Fall vollständig fehlender Werte verwandt worden: $\mathcal{Y}^* = \{\{y\} \,|\, y \in \mathcal{Y}\} \cup \{\mathcal{Y}\}$. Sei $R(u,.)$ eine Variable, die das Antwortverhalten von u bei der Frage nach dem Haushaltseinkommen darstellt, also $R(u,.) := 1$, wenn $Y(u,.)$ genau angegeben wird, $R(u,.) = 2$ sonst. Wenn nun die Existenz einer Variablen $R^*(u,.)$ mit $R(u,.) = 1 \Leftrightarrow R^*(u,.) \geq 0$ und ein Zusammenhang zwischen $Y(u,.)$ und $R^*(u,.)$ postuliert wird, dann ergibt sich für den Erwartungswert der Stichprobe mit vollständigen Antworten $\mathbb{E}(Y(u,.) \,|\, R^*(u,.) \geq 0)$. Dies muss nicht mit dem unkonditionalen Erwartungswert $\mathbb{E}(Y(u,.))$ übereinstimmen, wenn $R^*(u,.)$ und $Y(u,.)$ stochastisch abhängig sind. Andererseits erfordert die CAR-Bedingung bei fehlenden Werten die Gleichheit von $\mathbb{E}(Y(u,.))$ und dem bedingten Erwartungswert von $Y(u,.)$ gegeben $R(u,.) = 1$. Das Modell ist also sicher nicht CAR, falls ein stochastischer Zusammmenhang zwischen $R^*(u,.)$ und $Y(u,.)$ unterstellt wird. Jede gemeinsame Verteilung von $Y(u,.)$ und $R^*(u,.)$ erzeugt ein

Scharfstein und Irizarry 2003; van der Laan und Robins 2003). Beide Verteilungen sind aber nicht von direktem Interesse und ihre nichtparametrische Schätzung, die zusätzliche probabilistische Annahmen vermeidet, ist bei den üblichen Stichprobengrößen wie im ALLBUS sehr instabil.

Selektionsmodell, das nur dann ignorierbar ist, wenn $Y(u,.) \perp\!\!\!\perp R^*(u,.)$ gilt. Denn der Ausdruck

$$\begin{aligned}\Pr(Y^*(u,.) = \mathcal{Y} \mid Y(u,.) = y) &= \Pr(R(u,.) = 2 \mid Y(u,.) = y)\\ &= \int_{-\infty}^{0} f(v \mid Y(u,.) = y)\, dv\end{aligned}$$

müsste aufgrund von (4.1) konstant in y sein, wenn es sich um ein CAR-Modell handelte. Es ist aber von vornherein klar, dass über ein solches nicht ignorierbares Selektionsmodell empirisch wenig zu sagen sein wird, da weder $R^*(u,.)$ noch $Y(u,.)$ tatsächlich beobachtet werden.

Sind $Y(u,.)$ und $R^*(u,.)$ gemeinsam normalverteilt mit

$$\begin{aligned}&\mu \coloneqq \mathbb{E}(Y(u,.)), \quad \mu^* \coloneqq \mathbb{E}(R^*(u,.))\\ &\mathrm{Var}(R^*(u,.)) \coloneqq 1, \quad \sigma^2 \coloneqq \mathrm{Var}(Y(u,.))\\ &\rho \coloneqq \mathrm{Corr}(Y(u,.), R^*(u,.))\end{aligned}$$

dann ist die bedingte Dichte von $Y(u,.)$ gegeben $R(u,.) = 1$

$$\begin{aligned}\phi(y \mid R^*(u,.) \geq 0) = \frac{1}{\Pr(R(u,.) = 1)} \frac{1}{\sigma} \phi((y-\mu)/\sigma) \times\\ \Phi\left(\frac{1}{\sqrt{1-\rho^2}}(\mu^* + \frac{\rho}{\sigma}(y-\mu))\right)\end{aligned} \tag{4.9}$$

wobei ϕ bzw. Φ hier die Dichte bzw. Verteilungsfunktion der standardisierten Normalverteilung bezeichnen (Copas und Li 1997: 59f). Der letzte Term ist die bedingte Wahrscheinlichkeit für eine vollständige Angabe bei gegebenem Wert von $Y(u,.)$

$$\Pr(R = 1 \mid Y(u,.) = y) = \Phi\left(\frac{1}{\sqrt{1-\rho^2}}(\mu^* + \frac{\rho}{\sigma}(y-\mu))\right)$$

Diese bedingte Wahrscheinlichkeit entspricht der bedingten Wahrscheinlichkeit in (4.1). Die CAR-Bedingung gilt genau dann, wenn der Koeffizient von y 0 ist, wenn also $\rho = 0$ ist.[17]

[17] In der ökonometrischen Literatur wird Heckmans Modell oft als ein Beispiel für eine

Da $R(u,.)$ für alle $u \in s$ bekannt ist, kann wegen $\Phi(\mu^*) = 1 - \Phi(0;\mu^*) = \Pr(R^*(u,.) \geq 0) = \Pr(R(u,.) = 1)$ ein Schätzer von μ^* durch $\Phi^{-1}(\mathrm{P}(R = 1))$ konstruiert werden, wobei $\mathrm{P}(R = 1)$ den Anteil vollständiger Beobachtungen in der Stichprobe s angibt. Dagegen müssen μ, σ und ρ auf der Basis der vollständigen Beobachtungen geschätzt werden. Insbesondere der Parameter ρ, der die Abweichung von einem CAR-Modell darstellt, lässt sich nur aufgrund der Abweichung der Verteilung der vollständigen Beobachtungen von einer Normalverteilung identifizieren. Wenn im Fall des Haushaltseinkommens die genauen Angaben zur Berechnung der Parameter μ, σ, ρ verwandt werden, so erhält man die völlig unplausiblen Werte $\hat{\mu} = 1312$, $\hat{\sigma} = 3119$, $\hat{\rho} = 0,987$. Der geschätzte Erwartungswert liegt selbst außerhalb des konsistenten Mittelwertintervalls auf der Bruttostichprobe s_1 (vgl. Abschnitt 4.5). Nach diesem Ergebnis hätte ein erheblicher Teil der Population ein stark negatives Haushaltseinkommen. Zudem wäre die „Antwortwortbereitschaft“ $R^*(u,.)$ fast perfekt mit dem Einkommen korreliert. Benutzt man dagegen die logarithmierten Einkommen, ergibt sich als Erwartungswert 3849 DM und $\hat{\rho} = -0,845$, also ein stark negativer Zusammenhang zwischen Einkommen und „Antwortbereitschaft“. Etwas stabilere Ergebnisse wird man nur erhoffen können, wenn Kovariablen mit großem Effekt auf $R(u,.)$ angegeben werden können, die keinen Einfluss auf $Y(u,.)$ haben. Aber selbst dann wird die Identifikation des Modells im wesentlichen durch Linearitätsannahmen im Modell für $R^*(u,.)$ ermöglicht. Denn da es zu gegebenen Daten immer ein (konditionales) CAR-Modell gibt, kann der Koeffizient von y in der bedingten Wahrscheinlichkeit $\Pr(R(u,.) = 1 \mid Y(u,.) = y, X(u,.) = x)$ immer als 0 angenommen werden, wenn nur der Einfluss der übrigen Kovariablen $X(u,.)$ allgemein modelliert wird. Die Identifikation des Koeffizienten von y und damit der Korrelation ρ erfordert eine parametrische Einschränkung der Wirkung der Kovariablen. Aber die Auswirkungen etwa von Linearitätsannahmen in $\Pr(R(u,.) = 1 \mid Y(u,.) = y, X(u,.) = x)$ sind ebenso wenig wie die Auswirkungen willkürlicher Verteilungsannahmen

„selection on unobservables“ angeführt (Nicoletti 2002: 3). Die Bezeichnung soll wohl auf die Abhängigkeit des Modells von der unbeobachteten Variablen R^* hinweisen. Die Bezeichnung ist irreführend, denn wie die letzte Gleichung zeigt, hängt die Selektion nicht von unbeobachtbaren Größen ab, sondern von den Werten von $Y(u,.)$, etwa dem Haushaltseinkommen. Letzteres ist zwar nicht für alle Befragten $u \in s$ bekannt, aber es ist sicher nicht „unbeobachtbar“.

einfach zu überblicken.

Immerhin ergibt das Heckmansche Modell einen ersten Ansatzpunkt zur Modellierung nicht ignorierbarer Selektionsprozesse. Zudem können auch partielle Angaben relativ leicht einbezogen werden. In der Tat kann das Modell so erweitert werden, dass viele häufig auftretende Probleme mit unvollständigen Angaben in diesem Rahmen formuliert werden können (z.B. Crouchley und Ganjali 2002). Weiterhin wurden Verteilungsannahmen abgeschwächt (Das et al. 2003) und asymptotische Analysen verfeinert (Rotnitzky et al. 2000). Ausgehend von Heckmanschen Selektionsmodellen können die Auswirkungen „lokaler" Abweichungen von den Annahmen approximiert werden (Copas und Li 1997).

4.11. Sensitivität

In der sozialwissenschaftlichen Praxis reicht es zumeist nicht, sich für ein Selektionsmodell zu entscheiden und sodann Ergebnisse nur unter dieser Modellannahme zu präsentieren. Wird mit einer CAR-Annahme begonnen, so muss dennoch Rechenschaft über die Konsequenzen der Annahme abgelegt werden. Das ist umso wichtiger, als die CAR-Annahme dazu verführen könnte, nicht über Selektionsprozesse nachzudenken. Zwar können „lokale" Abweichungen von der CAR-Annahme allgemein beschrieben werden. Damit ist es auch möglich, die Konsequenzen der CAR-Annahme für statistische Aussagen zu approxmieren (Copas und Eguchi 2001). Aber probabilistische CAR-Modelle können kaum als realistische Modelle für Stichprobenausfälle angesehen werden. Daher wird eine lokale Approximation, so nützlich sie theoretisch ist, zumeist nicht befriedigen können. Die Approximation läuft Gefahr, die Konsequenzen von Annahmen über den Selektionsprozess nur in den engen Grenzen probabilistisch ähnlicher Modelle zu untersuchen. Rosenbaums Sensitivitätsanalyse (2002: Chap. 4) knüpft an Überlegungen über die Variable $R^*(u, .)$ an, wie sie auch in Heckmans Modell verwandt wird. Rosenbaums Methode geht von einer größeren Menge von Selektionsmodellen aus und erlaubt damit eine umfassendere Abschätzung der Effekte von nicht ignorierbaren Selektionsprozessen. Die Methode verbleibt aber in der unterstellten Modellwelt und ist zudem bisher nur für spezielle Statistiken formuliert worden.

Robins und Mitarbeiter haben vorgeschlagen, von den gewichteten Schätzfunktionen

$$\sum_{u \in s} \frac{\mathbb{1}[R(u,.) = 1]}{\pi(u, X)} (Y(u,.) - \mu)$$

auszugehen und den Propensity Score auch als Funktion von $Y(u,.)$ zu modellieren (Rotnitzky und Robins 1997; Robins 1997; Rotnitzky et al. 1998; Scharfstein et al. 1999; Scharfstein und Irizarry 2003). Da

$$\begin{aligned}
&\mathbb{E}\left(1 - \frac{\mathbb{1}[R(u,.) = 1]}{\pi(u, X, Y)}\right) \\
&= 1 - \mathbb{E}\left(\frac{1}{\pi(u, X, Y)} \mathbb{E}\left(\mathbb{1}[R(u,.) = 1] \mid X(u,.), Y(u,.)\right)\right) = 0
\end{aligned}$$

ist, kann die Schätzfunktion erweitert werden:

$$\sum_{u \in s} \frac{\mathbb{1}[R(u,.) = 1]}{\pi(u, X, Y)} (Y(u,.) - \mu) + \left(1 - \frac{\mathbb{1}[R(u,.) = 1]}{\pi(u, X, Y)}\right) \phi(X; \mu)$$

Wird etwa logit $\pi(u, x, y) = x(u)\beta + \alpha y(u)$ gesetzt und

$$\phi(X; \mu, \alpha) = \frac{\mathbb{E}(Y(u,.) \exp(\alpha Y(u,.)) \mid R(u,.) = 1, X(u,.))}{\mathbb{E}(\exp(\alpha Y(u,.)) \mid R(u,.) = 1, X(u,.))} - \mu$$

gewählt, dann ergibt sich ein *doppelt robuster* Schätzer, der auch dann konsistent ist, wenn die Auswahlgleichung fehlspezifiziert ist. Dabei kann der Koeffizient α von $Y(u,.)$ in der Auswahlgleichung $\pi(u, X, Y)$ nur aufgrund der Beobachtungen mit vollständigen Angaben zu $Y(u,.)$ geschätzt werden, wenn also $R(u,.) = 1$ ist. Da außerdem zu jedem Selektionsmodell auch ein (konditionales) CAR-Modell gebildet werden kann, wenn nur $\pi(u, X)$ flexibel genug gewählt wird, ist α auch in sehr großen Datensätzen kaum stabil schätzbar. Robins und Mitarbeiter haben daher vorgeschlagen, eine Reihe von Werten des Koeffizienten α fest zu wählen und die Auswirkungen auf die Schätzung von μ zu notieren. Wie ihre Simulationen und Beispiele aber zeigen, sind auch diese Schätzungen sehr instabil. Zudem erscheint noch unklar, wie partielle Angaben in diesem Ansatz direkt berücksichtigt werden können (Robins 1997; Robins und Gill 1997).

Manskis Abschätzungen der Folgen von Selektivität (Manski 1993; Manski und Horowitz 2000; Manski 2003; Zaffalon 2002; Manski und Tamer 2002) erfolgen ähnlich wie die Abschätzungen in Abschnitt 4.5 und verzichten auf probabilistische Annahmen über den Selektionsprozess. Sie führen nur dann zu relativ engen Intervallen, wenn $Y(u,.)$ beschränkt ist. Im Fall des Haushaltseinkommens ergeben sich daher Intervalle für die möglichen Werte der Durchschnitte wie schon im Abschnitt 4.5. Diese Intervalle sind auch die theoretischen Grenzen der Methode von Robins et al., wenn ihr Koeffizient α alle Werte zwischen $-\infty$ und ∞ durchläuft (Scharfstein et al. 1999: 1108). Raghhunathan hat in der Diskussion der Arbeit von Manski und Horowitz (2000: 86) die Weite der Intervalle beklagt: "‘I am afraid that I agree with Cochran (1977) that such an approach is so conservative as to be of little value in most practical settings for inferential purposes."’ Manski und Horowitz antworteten:

> The width of the bounds reflects the information available from the data per se about the population parameters of interest. The width also indicates the relative importance of the data and untestable assumptions in determining the values of point estimates. ...Readers should be told that the point estimates are sensitive to untestable assumptions and that different assumptions could produce widely different results. (Manski und Horowitz 2000: 87f).

4.12. Diskussion

Soziologen haben bisher die Möglichkeiten und vor allem die Grenzen von Selektionsmodellen selten zur Kenntnis genommen. Zwar gab es immer wieder Arbeiten in verbreiteten Zeitschriften, die das Thema aufgegriffen haben (Berk und Ray 1982; Stolzenberg und Relles 1997; Winship und Mare 1992). Aber in empirischen Arbeiten wird das Problem oft vollständig ignoriert, mit einem kurzen Verweis auf MAR-Modelle abgetan oder mit einem Heckman Modell erledigt. Bei Ausschöpfungsraten von 50% in Umfragen wird man es sich nicht auf Dauer leisten können, Selektionsmodelle, ihre Grundlagen und ihre Konsequenzen zu ignorieren.

Nun kann über die Entstehung fehlender Angaben immer nur spekuliert werden. Man ist auf Informationen einschlägiger Untersuchungen über Antwortverhalten angwiesen. Existieren aber gruppierte, zensierte oder fehlklassifizierte Berichte der Befragten, dann müssen diese Angaben ernst genommen und in die Berechnung von Statistiken einbezogen werden. Es werden Verfahren benötigt, die auch diese partiellen Angaben berücksichtigen. Selektionsmodelle erlauben es, über die Voraussetzungen und Konsequenzen solcher Verfahren nachzudenken. Zudem sind in diesem Fall Spekulationen über Selektionsprozesse empirische Grenzen gesetzt und Selektionsmodelle können auch ohne Rückgriff auf zusätzliche Informationen beurteilt werden. Dazu können Methoden der Sensitivitätsanalyse einen wesentlichen Beitrag leisten. In dieser Rolle werden Selektionsmodelle auch für die empirische Sozialforschung an Bedeutung gewinnen.

4.13. Postscriptum

Selection models have seen a tremendous development in recent years, including the construction of new models, clever estimation strategies and theoretical analyses. On a theoretical level, the interplay between parametric specifications, model completeness, sufficient statistics, and incomplete data has been clarified by Lu and Copas (2004). Furthermore, the algorithmic description of CAR models has been completed by Gill and Grünwald (2005).

The practical methods developed concentrate on imputation, propensity scores and matching methods, and inverse probability weighted estimators. Recent developments of imputation methodologies are reviewed by Tang et al. (2005), Zhang (2003), Harel and Zhou (2007), Horton and Lipsitz (2001), Kenward and Carpenter (2007), Chen et al. (2000), and

Ambler et al. (2007). The role of proper imputation rules is discussed by Nielsen (2003a) and Shafer (2003). The possible role of imputation procedures for non-ignorable missing and incomplete data cases is surveyed by Demirtas (2005). Asymptotic and finite sample properties were further examined by Kim (2004). The possible use of imputations for model checking (Gelman et al. 2005 and Abayomi et al. 2008), disclosure control (Little et al. 2004), in binary data models (Demirtas and Hedeker 2007 and Münnich and Rässler 2005) and in structural models (Song and Belin 2004) were explored as well as more special aspects such as the bootstrap (Zhang 2004), sensitivity analysis (Carpenter et al. 2007), local procedures (Aerts 2002) and the effects of outliers (Elliott and Stettler (2007) and of rounding (Horton et al. 2003). There has also been much interest in applications in survey sampling (Beaumont 2005, Haziza and Rao 2005, Kim et al. 2006, Scheuren 2005, Skinner and Rao 2002, and Shao and Wang 2008). There have also been developments of conditional methods (van Buuren 2007), of empirical likelihood methods (Wang and Rao 2002), nearest neighbour methods (Wasito and Mirkin 2006) and of adaptive procedures (Reiter and Raghunathan 2007). There are by now many discussions of practical procedures in the context of well known data sets (Nicoletti 2006 discusses the ECHP, Frick and Grabka 2003 the SOEP). Moreover, there has been some interest of applying imputation techniques to survival data (Rubin and van der Laan 2005 and Hsu et al. 2006, 2007). Comparisons to other methods are provided by Kim and Park (2006) and Carpenter et al. (2006).

The propensity score based procedures are discussed in general by Lunceford and Davidian (2004), D'Agostino (2004), Austin (2008), Frölich (2004, 2005, 2007), and Senn et al. (2007). An interesting discussion of practical implementations and problems of inference is provided by Dehejia and Wahba (2002), Smith and Todd (2001, 2005a, 2005b), and Dehejia and Wahba (2005). The balancing effect of propensity score matching procedures is investigated by Austin et al. (2007b) and Berger (2005a). Variable selection is discussed by Judkins et al. (2007) and Brookhart et al. (2006). Possible bias arising from propensity score procedures is discussed by Austin et al. (2007a). The handling of multivalued and continuous variables is presented by Hirano and Imbens (2004) and Imai and van Dyk (2004). The connection with instrumental variables is taken up by Ichimura and Taber (2001). Kurt et al. discuss applications under non-standard conditions, Lu (2005) presents results about the use of propensity scores with time dependent covariates and McCaffrey et al. (2004) present first methods of boosted regression used in conjunction with propensity score techniques. Finally, Leon and Hedeker (2007) present results on using misspecified propensity scores. The latter problem is closely connected to the discussion of doubly robust estimators as developed by Robins and co-workers. A good recent review of constructions and properties of doubly robust estimators is given by Kang and Schafer (2007) and its discussion. The arguments of Neugebauer and van der Laan (2005) and van der Laan and Robins (2003) on the practical advantages of doubly robust procedures are also relevant here.

Methods of empirical likelihood for incomplete data models have been developed by Qin and Zhang (2007, 2008), Chen and Qin (2006), Leung and Qin (2006), and Wang and Rao (2002). General methods for sensitivity analyses are provided by Hines and Hines (2006, 2007) and Frank and Min (2007).

Progress for interval methods as surveyed by Manski (2003) is rather slow. Horowith and Manski (2006) discussed general identification problems. Dantsin et al. (2006) use interval methods for incomplete data to provide an algorithm for the variance if the interval lengths are small in comparison to the distances between the interval centers. This relates to the results discussed briefly in Section 4.5. Imbens and Manski (2004) suggest confidence bands for interval estimates, but their mix of fixed sample and probabilistic concepts is not too convincing. Blundell et al. (2007) give a very readable application of general interval methodology for the analysis of wage differentials. Scharfstein et al. (2004) provide another example.

There is by now a rather specialised literature treating missing and incomplete covariates in regression models and in censored regression models in particular. Recent contributions include Catchpole et al. (2008), Chen (2002), Chen and Little (1999, 2001), Chen et al. (2007), Didelez (2002), Dupuy and Mesbah (2004), Dupuy et al. (2006), Herring et al. (2004), Huang et al. (2005), Lee and Tang (2006), Liang (2008), Liang et al.

(2004), Lipsitz et al. (2004), Mojirsheibani and Montazeri (2007), Nielsen (2003b), Paliwal and Gelfand (2006), Parzen et al. (2006), Pons (2002), Qi et al. (2005), Rathbun et al. (2007), Rathouz (2003, 2007), Roy and Lin (2005), Shardell and Miller (2008), Stubbendick and Ibrahim (2003, 2006), Tian and Lagakos (2006), Wang and Paik (2006), Wang and Chen (2001), Wang et al. (2007), Wang et al. (2001), Wang and Wang (2001), Wang (2005), Wu (2004, 2007, 2008), Yang et al. (2005), and Zhang and Rockette (2005, 2006, 2007). Covariates missing by design are treated by D'Angelo and Weissfeld (2007), intermittendly missing covariates in longitudinal studies are discussed by Andersen and Liestøl (2003), Gad and Ahmed (2006, 2007), and Lin et al. (2004). Censored covariates are discussed by Dabrovska (1995), Yashin and Manton (1997). Missing censoring indicators, a form of missing information often encountered in recurrent event data, are discussed by Wang and Shen (2008), Antony and Sankaran (2008), Subramanian (2004, 2006), and Lu and Liang (2008).

5

Causal Inference from Series of Events

5.1. Summary

Recent years have witnessed an increased interest, both in statistics and in the social sciences, in time dependent models as a vehicle for the causal interpretation of series of events. The Humean and empiricist tradition in the philosophy of science uses the constant temporal order of cause and effect as a decisive delimitation of causal processes from mere coincidences. To mimic the philosophical distinction, series of events are modelled as dynamic stochastic processes and the precedence of cause over effect is expressed through conditional expectations given the history of the process and the history of the causes. A main technical tool in this development is the concept of conditional independence.

In this article we examine some difficulties in the application of the approach within empirical social research. Specifically, the role of probabilistic concepts of causality and of conditional independence, the nature of events that reasonably qualify as causes or effects, and the time order used in empirical research are considered.

5.2. The Tradition of Causal Analysis in Sociology

The use of the concept of causality in sociology has been lingering between neglect and over–reliance.[1] Even though the concept was never wholeheartedly accepted by sociologists, it became a cornerstone of arguments in favour of empirical research by the early 1970's. The work of Lazarsfeld, Blalock, Coleman, Duncan, and many others led to the predominance of path models as the paradigmatic form of statistical analyses of causation, which was conceived as a relation between statistical variables. The flowering of structural equation modelling within sociology and psychology strengthened the technical applicability of the approach and gave impetus to the development of statistics in general.[2] But the increased statistical sophistication was accompanied by a growing isolation from the rest of applied social research as well as from statistics (as a branch of applied mathematics). Critical discussions within sociology and across disciplinary boundaries were prematurely cut off. Even the debates in econometrics during the late 50's and early 60's[3] on 'autonomy' of causes, the meaning of simultaneous equations, and the concept of exogeneity are rarely reflected in the textbooks on social statistics from the late 70's onwards. Moreover, developments in statistics, even when originating from concerns for questions of causality from other disciplines, were largely ignored.[4] And within sociology, the connection of causal analysis with certain statistical techniques was met by a general scepticism concerning the role of statistics and of 'variables' in general.

Sociologists outside the tradition of structural equation models often downplay the role of causality in the social sciences in favour of other forms of determination. In fact, the sociological literature abounds with examples of explanations that are not strictly causal in an empiricist or positivistic sense. Historical, functional, structural, teleological explanations—to name just a few of the distinctions used—are often

[1] See Bernert (1983) for an account of its history in the American sociological literature.

[2] See Clogg (1992) for a partial review, including other areas of social statistics.

[3] See e.g. Epstein (1987).

[4] Holland's discussion of Clogg's 1992 paper and Clogg's rejoinder may serve as an illustration.

invoked. As far as causal reasoning is granted a place in sociology, many researchers agree with a view of causality that depends on subject-matter considerations. There seems to be wide consensus among sociologists that causality cannot simply and directly be inferred from empirical data, regardless of whether they are obtained from randomised experiments, collected through ingenious research designs or summarised by particularly advanced statistical models. Blumer's well known early (1956) diatribe against statistical 'variable analysis' is but one example in the sociological tradition asking for more than a statistical analysis of the relation between dependent and independent variables.

More recently, however, statisticians, sociologists, and philosophers have begun to study the relation between statistics and causality more closely. The renewed interest was sparked by advances in the formalisation of concepts related to causation. Most of these developments are well documented in a special volume of Synthese (vol. 121, 1999) and —with emphasis on social science applications— in the proceedings "Causality in Crisis", edited by McKim and Turner (1997). Most of the contributions in these recent publications concentrate on counterfactual or interventionist conceptions of causation. Complementary, we will here investigate the prospects of a classical empiricist criterion of causality in the Humean tradition: the time order of cause and effect.

5.3. Temporal and Probabilistic Criteria of Causation

Since many theories of causal interdependence rely on empiricist criteria as a prerequisite for the acknowledgement of causality, a strong argument against the assumption of a causal connection can be made if some of the empiricist criteria of causality do not hold. The empiricist conditions for the existence of a causal relation, based on Hume's analysis, require a) spatial and temporal contiguity, b) constant conjunction between cause and effect, and c) temporal succession. We will argue that suitably modified versions of the requirements b) and c) can be used as starting points for empirical arguments concerning causal claims.

The first criterion, that of spatial and/or temporal contiguity, is often disputed. Effects do not need to follow immediately after a cause, nor do they need to be spatially close to the cause. E.g. strikes and demonstrations may be the (efficient) cause of a change of government, but the latter need not follow immediately after the demonstrations, nor need there be any spatial contiguity. Even though criterion a) cannot generally serve as a necessary condition for causation in the social sciences, a formalism of cause–effect relationships should be able to distinguish between 'close' and 'remote' causes of effects. Otherwise it would be difficult to express the ideas of 'spurious cause' and 'causal chain', both considered useful in the construction and criticism of causal explanations.

Condition b) cannot be expected to hold strictly with respect to social interactions. People tend to react differently upon others actions, even in otherwise similar situations. And, as Suppes (1970: 92) notes:

> Empirical studies of the sort done in psychology, sociology, and medicine hold little hope of establishing complete deterministic chains for the causes of actions. This is true whether we are analysing the sex habits of Eskimos, recidivism among parolees over forty, or church attendance by illiterates.

The 'constant conjunction' condition therefore must be reformulated. Here we will simply replace the 'constant conjunction' condition by a probabilistic relation: that the cause changes the probability of the effect.[5] Note that "[i]t is not a matter of presenting evidence for causality by offering probabilistic considerations but it is part of the concept itself to claim relative frequency of co-occurrence of cause and effect" (Suppes 1970: 45). The incorporation of probabilistic aspects is thus not only of an epistemic nature. It is not incorporated because the sociologist does not (yet) know, but because 'constant conjunction' cannot reasonably

[5] Suppes' (1970) analysis starts from the assumption that causes should *increase* the probability of the effect. The subsequent discussion in the philosohical literature has shown that the concept of 'increase of probability' may not suffice for the analysis of probabilistic causality (e.g. Eells 1991). But since the problem is not central to our discussion we will content ourself with the simple minded principle of a change of probability.

be claimed in sociology. Consequently, probability statements in this context should generally refer either to frequencies or to propensities, as also fits the needs of an empiricist program.

The condition c) requires a temporal framework for causal arguments. This is not normally included in the formulation of 'causal models' in the structural model literature. It is also not included in the formal representation of observational studies[6] in e.g. biometry nor in more recent counterfactual or interventionist accounts. On the other hand, the temporal dimension of causal connections has often been recognised in sociology. Tuma and Hannan, introducing event–history analysis into sociology, acknowledge the interplay between temporal and causal analysis (1984: xi–xii, their italics):

> Any attempt at forging a systematic framework for the empirical study of social change must confront two issues. One involves the development of *dynamic* models—models that describe the time paths of change in phenomena. The other involves the development of *causal* models—models that describe how change in some properties induces change in still other properties. …we rely heavily on the use of formal models to guide attempts at testing hypotheses about the processes and causes of change.

And Blossfeld and Rohwer (1995: 20) state:

> …the important task of event history modelling is …to establish relevant empirical evidence that can serve as a link in a chain of reasoning about causal mechanisms. In this respect, event history models might be particularly helpful instruments because they allow a time-related empirical representation of the structure of causal arguments.

[6] The definition of an observational study is a rather narrow one in this context. Rosenbaum (1995: 1) states: "An observational study concerns treatments, interventions, or policies and the effect they cause, and in this respect it resembles an experiment. A study without a treatment is neither an experiment nor an observational study." Following his definition, many empirical sociological studies will not count as observational studies.

Suppes (1970) proposed one of the best known formalisations of probabilistic causality. His account includes a condition of temporal precedence of cause over effect, in contrast to many later attempts to clarify the concept (e.g. Eells, 1991). His starting point is the theory of probability based on systems of sets, called 'events'. He adjoins a temporal indicator to these sets to indicate temporal sequences. But as witnessed by later contributions (e.g. Davis 1988), this strategy turned out to be rather limited. Savage (1972: 10) remarked with respect to the use of the term 'event' in probability theory: "...the concept of event as here formulated is timeless, though temporal ideas may be employed in the description of particular events." Arjas and Eerola (1993: 384) note that in these formulations, "time is present (if at all) only as an index, distinguishing between what comes 'before' and what comes 'after'." Consequently, a more flexible representation of time order is needed. In the absence of condition a), it should at least be possible to formulate the timing of effects with respect to causes. Recent contributions therefore seek to enrich the formulations by borrowing heavily from dynamic theories of stochastic processes.

The program then is clear: to combine probabilistic models —re-expressing the constant conjunction (condition b)— with the idea of 'the cause precedes the effect' (condition c) to facilitate an empirical assessment of claims of causality. Causal statements are translated into the mathematical language of stochastic processes: $Y = \{Y_t, t \in \mathcal{T}\}$ is a stochastic process with values in a finite set $\mathcal{Y}$. The values of Y_t are interpreted as properties of *units* under study and *events* are changes of properties at time points t. At any given point in time t, the description of the evolution of the process may depend on conditions and events that occurred in the past, i.e. before t, but not on what is the case at t or in the future, after t. Causes acting on Y may then be introduced by considering them as changes in a further process $X = \{X_t, t \in \mathcal{T}\}$. The process X may be treated as a time-dependent covariate and causal statements are therefore formulated as probabilistic relations between two (or more) stochastic processes. A time-dependent covariate records when a causal factor has changed its state. It signifies that an event of kind $\mathcal{X}$ has taken place. Consequently, we would not say that a process X is a cause of a process Y, but that a change in X at time t (an event at time t) could be a cause of a change in Y at time t', $t' > t$ (another event at a later time).

Often a canonical dynamic description of the stochastic processes can be given, one relating the 'past' of the processes to their 'future', and encapsulating their relevant probabilistic features. In this case the classical formulation of probabilistic causality (e.g. Suppes 1970) can be enhanced by allowing for an explicit representation of the timing of effects, generalising the Humean requirement of contiguity in time. Approaches along this line were advocated by Granger (1969) in the context of time series analysis and by Schweder (1970, 1986) for general Markov processes. These ideas were taken up more recently by Aalen (1987), Arjas/Eerola (1993), Parner/Arjas (1999), and Keiding (1999), providing a formal framework for probabilistic causality with a clear relation to time order. Many of these articles use genuinely epistemic notions for the interpretation of probabilities. This partly reflects the naming conventions used in the technical literature on stochastic processes and we will follow the convention here. But the mathematical formulation does not force us to accept an epistemic interpretation, and the possibility of interpretations in terms of propensities or frequencies should be kept in mind.

5.4. Mathematical Models

A dynamic description of stochastic processes fitting the above program can be outlined in the case of processes with discrete time parameters:[7]

[7] Much of the discrete-time theory extends directly to the continuous-time setting. The reason is that a continuous time martingale with respect to a right continuous filtration can be modified to have nice sample path properties, i.e. right-continuous paths with left hand limits. A thorough treatment presupposes a formidable technical machinery without adding much insight to the present discussion. But it should be noted that the continuous-time theory makes heavy use of continuity and of the denseness of the rationals within the real numbers. Many mathematical models of time try to avoid such strong assumptions (see Whitrow 1963: Chap. III for an early review). Moreover, the combination of a continuous-time theory using all the properties of the reals may collide with a concept of causality that is based both on the time ordering of cause and effect and on the distinction between direct and indirect causes (see e.g. Suppes 1970: 72).

Let $Y = \{Y_t, t = 0, 1, 2, \ldots\}$ be a stochastic process with values in a finite set $\mathcal{Y}$.[8] Suppose one is interested in what happens just after time $t - 1$. A good prediction is the conditional expectation of the change of Y_t from Y_{t-1}, conditional on the previous history of the process, that is the random variable

$$V_t = \mathbb{E}(Y_t - Y_{t-1} \mid Y_0, \ldots, Y_{t-1}) \tag{5.1}$$

Putting

$$U_t = \sum_{s=1}^{t} V_s = \sum_{s=1}^{t} \mathbb{E}(Y_s - Y_{s-1} \mid Y_0, \ldots, Y_{s-1})$$

for the sum of the predicted values, one may write the original process Y_t as a sum of predictions in time and a remainder, the Doob decomposition:[9]

$$Y_t = Y_0 + U_t + M_t \tag{5.2}$$

The prediction part U_t is a function of the previous history up to and including time $t - 1$ only, while M_t is a *martingale*, satisfying

$$\mathbb{E}(M_t \mid Y_0, \ldots, Y_{t-1}) = M_{t-1} \tag{5.3}$$

In fact, for the *martingale difference* $M_t - M_{t-1}$ one finds

$$\begin{aligned}
&\mathbb{E}(M_t - M_{t-1} \mid Y_0, \ldots, Y_{t-1}) \\
&= \mathbb{E}((Y_t - Y_{t-1}) - (U_t - U_{t-1}) \mid Y_0, \ldots, Y_{t-1}) \\
&\qquad = \mathbb{E}(Y_t - Y_{t-1} \mid Y_0, \ldots, Y_{t-1}) - V_t = V_t - V_t = 0
\end{aligned}$$

The decomposition (5.2) generalises the additive regression decomposition into a 'structural' part U_t and a 'random' part M_t that is used in much of empirical social research. Accordingly, but somewhat ambiguously, the differences $M_t - M_{t-1}$ are sometimes called the *innovations*

[8] The assumption of a finite state space is not essential for the mathematical formulations used here. But subsequent discussions of the appropriateness of the formal model often presuppose a finite number of states.

[9] Williams (1991) provides a thorough and vivid exposition for the discrete time case.

of the process. The predictions U_t, depending only on $Y_{t-1}, \dots, Y_0$, are called *predictable*. Any other process Z_t that depends only on the values of $Y_{t-1}, \dots, Y_0$, the history of Y strictly before t, will also be called *predictable* with respect to the process Y.

It may be instructive to see how a duration variable T, featuring prominently in event-history analysis, fits into the present framework. In that case one may put $Y_t = 1$ if the event happened at time t or before, 0 otherwise. That is, $Y_t = \mathbb{1}[T \leq t]$, where $\mathbb{1}[A]$ is the indicator variable of the event A. For simplicity, one may also assume $Y_0 = 0$. The prediction process is then given by $V_t = \mathbb{E}(Y_t - Y_{t-1} \mid Y_0, \dots Y_{t-1})$. This quantity is 0 if Y_{t-1} takes the value 1, since then both Y_t and Y_{t-1} must be 1. In other words: If the event happened before time t, there will be no change in the prediction of Y_t, because the one possible change in state is known to have occurred.

On the other hand, if there was no event at $t - 1$ or before, Y_{t-1} as well as $Y_{t-2}, \dots, Y_0$ are 0 and V_t reduces to the probability $\Pr(Y_t = 1 \mid Y_0 = 0, \dots, Y_{t-1} = 0) = \Pr(T = t \mid T \geq t)$. Thus V_t reduces to a random function of the well known hazard function of T. If $\lambda(t) = \Pr(T = t \mid T \geq t)$ denotes the hazard function of T, $V_t = \mathbb{1}[T \geq t]\lambda(t) = (1 - Y_{t-1})\lambda(t)$. Note that V_t in the present discussion is a random variable, depending on $Y_{t-1}, \dots, Y_0$. In contrast, the hazard function $\lambda(t)$ treated in most texts on event–history analysis is a non-random transform of the distribution function. Furthermore, the accumulated prediction U_t is a random sum of hazard functions, the sum extending over all $s \leq \min(T, t)$, the times s before the event time T or the observation time t, whatever comes first.

Following Aalen (1987), a dynamic statistical model is then defined as a parameterisation of the prediction increments V_t. Since V_t depends on the history of the process up to and including $t - 1$ only, it can be interpreted as a description of the future, the likely events at time t, depending only on the knowledge of all past events. Alternatively, it is (an approximation of) the relative frequency of events at t among all sequences with this history.

To introduce concepts of interdependence between several processes, it seems natural to embed the above concepts into a multivariate extension.

Basically, the history of the single process is replaced by one based on all relevant information available before time t. In the case of two processes (Y_t, X_t) one might therefore define the conditionally expected increments as

$$V_t^Y = \mathbb{E}(Y_t - Y_{t-1} \mid Y_0, X_0, \ldots, Y_{t-1}, X_{t-1}) \tag{5.4}$$

and

$$V_t^X = \mathbb{E}(X_t - X_{t-1} \mid Y_0, X_0, \ldots, Y_{t-1}, X_{t-1}). \tag{5.5}$$

Interpreting the conditional expectations above as an increase in knowledge, V_t^Y will be based not only on the pre-t history of Y itself, but also on the knowledge of the development of X up to and including $t-1$. Thus, the expectation will change depending on the information provided by X. Symmetrically, V_t^X, the prediction of X based on the common history of X and Y before t, will depend on the changes in Y up to t. One may represent the two processes using the Doob decomposition with respect to the joint history of the processes as:

$$Y_t = Y_0 + \sum_{s=1}^{t} V_s^Y + M_t^Y \quad \text{and } X_t = X_0 + \sum_{s=1}^{t} V_s^X + M_t^X \tag{5.6}$$

Then X_t is defined not to be *causal for Y_t in Aalen's sense* if and only if, first, the prediction errors M_t^X and M_t^Y are uncorrelated, and second, U_t^Y, the prediction of Y_t, may depend on $Y_0, \ldots, Y_{t-1}$ but not on $X_0, \ldots, X_{t-1}$. We will say that X and Y are *locally autonomous* if the first condition is satisfied.[10] If the second condition holds, Y_t is said to be *locally independent* of X_t. Otherwise, Y_t is said to be *locally dependent* of X_t.[11]

[10] The notion of 'autonomy' has a long tradition in econometrics, especially in the context of simultaneous equation models. Aldrich (1989) provides a review.

[11] A related concept is Granger non-causality, a concept often used within econometric time series analysis. A process X is said not to cause Y in Granger's sense at t iff $Y_t \perp (X_{t-1}, \ldots, X_0) \mid Y_{t-1}, \ldots, Y_0$, i.e. where Y_t and the pre-t history of X, $(X_{t-1}, \ldots, X_0)$ are conditionally independent given the pre-t history of Y_t alone. Note that in the context of time series analysis the processes Y_t and X_t need not refer to 'events'. The concept has been explored and extended in a series of papers by Florens and Mouchart (1982, 1985).

The condition of local (in-)dependence is asymmetric in the two processes. Indeed, in the example of two duration variables $Y_t = \mathbb{1}[T_1 \leq t]$ and $X_t = \mathbb{1}[T_2 \leq t]$, the process X_{t-1} may enter as a time dependent covariate in the prediction (stochastic hazard) of the other, but not the other way around. This is in accord with the basic asymmetry of causal relations and also respects the notion of 'cause precedes effect'.

The condition of local autonomy is introduced to ensure a certain autonomy of the two processes. When satisfied it is possible to envisage a change in the behaviour of one process after time t without a change in its local relation to the other process or a corresponding immediate change in the other process. This should rule out processes that are merely related by definitions or 'rules of the game'. Consider for example two gamblers throwing dice. Denote by X^1 (X^2) the result of a throw of the first (second) player. If the throws are independent, then also the prediction errors are independent and X^1 and X^2 are autonomous. But one may also look at the result of the play, a win, a draw, or a loss, for each player. Let $Y^1 = \mathbb{1}[X^1 > X^2] - \mathbb{1}[X^2 > X^1] \in \{-1, 0, 1\}$. Then $Y^2 = -Y^1$ and the processes are clearly not autonomous.

5.5. Statistical Methods

An empirical strategy to show local dependence of Y on X at t is then to show that the prediction process U_t^Y with respect to the joint history of the process is a non-constant function of $X_{t-1}, \ldots, X_0$. In the case of simple duration models, $V_t^Y = \mathbb{1}[Y \geq t]\lambda_\theta(t; X_{t-1}, \ldots, X_0)$. That is, the process X appears as a time-dependent covariate in the hazard function for Y at t, which is assumed to be parameterised by some $\theta \in \Theta$. One therefore needs to show that the pre-t history of X changes the hazard of Y at t. Often it is possible to factor the likelihood of $(X, Y)_t$ for θ in such a way that only the part $U_t^Y(Y_t \mid Y_{t-1}, X_{t-1}, \ldots, Y_0, X_0; \theta)$ figures in the computation of statistics. This is called a *partial likelihood* for θ since it does not depend on the specification of the joint distribution of $(Y, X)_t$. In particular, a model for the covariate process X need not be specified. This allows for an attractive strategy to demonstrate local dependence since one can concentrate on the model for the conditional prediction of Y given X without worrying about the possibly complicated nature of X. In the context of counting process methods, Slud noted (1992: 97):

> ... that inferences could be made successfully without parametric specification of any probabilistic objects other than the failure counting process intensities. The latter point of view is especially liberating in problems with randomly time-varying covariates, where one is usually interested only in the effect of the covariates on the hazard of failure and where one can usually not provide convincing models of the stochastic variation of the covariates over time.[12]

5.6. Events and their Descriptions

The formal frame of causal reasoning developed above has many merits with regard to dynamic formulations of causality. Despite some conceptual shortcomings in the reformulation of the 'constant conjunction'

[12] Arjas/Haara (1984), Slud (1992) and Greenwood/Wefelmeyer (1998) have studied the statistical properties of factorisations with time-dependent covariates.

condition it may readily be used to formulate claims of (non-) causality. As it stands, it relates to stochastic processes, i.e. collections of random variables. Often, statisticians are satisfied with formal references to variables as causes and effects, especially when arguing in the tradition of structural equation models. E.g. Pearl and Verma (1992: 91) speak of "stable causal mechanisms, which on a microscopic level, are deterministic functional relationships between variables, some of which are unobservable." But neither variables nor what they stand for are generally admitted as causes or effects by social scientists. Background variables like sex or religion are not considered to be (representations of) possible causes.

One further prerequisite for the applicability of the above formalism in the social sciences is therefore a restriction on the entities that may be causes or effects. As Bunge (1963: 72, his italics) puts it, "*there can be no causal links among states*, nor among any other systems of qualities. States are not *causes*, but simply *antecedents* of later states. To regard states as causes amounts consequently to committing the fallacy of the *post hoc, propter hoc*." Thus neither states nor things nor qualities of things can be causes or effects, only *events* can.[13]

But what are events? Hacker (1982: 17) says that "[e]vents, unlike objects, are directly related to time. They occur before, after, or simultaneously with other events. They may be sudden, brief or prolonged…None of these temporal predicates apply in the same way to objects." But this special connection with time implies a difficulty for probabilistic theories of causality: "it is manifest that no event ever happens more than once, so that the causes and effects cannot be the same in *all* respects."[14] Therefore one cannot speak about the constant conjunction of cause and effect unless it is possible to also speak of *kinds of events*. While an event is something unique, events of the same kind can occur several times. But how can one define kinds of events? One possibility would be to delete some temporal descriptions from propositions about events. Such propositions might then be said to be about kinds of events.[15] The

[13] But see e.g. Mellor (1995: 129), who argues that "causation mostly links facts.… So no causation would be lost even if there were no particular events."

[14] Maxwell, cited in Bunge 1963: 50

[15] See e.g. Scheffler 1994.

obvious shortcoming of this approach is that it concerns propositions, not events, and propositions do not qualify as causes, at least not in a realist account of causality.

> Events presumably are not linguistic entities; like trees and molecules, events can be talked about, referred to, and described but they are not themselves statements, sentences, descriptions, or any other kind of linguistic units. Nor are events propositions; propositions are supposed to be abstract entities, whereas events are spatio-temporally bounded particulars. (Kim 1969: 198.)

An alternative is to relate events to changes in things. An *event*

> is a 'movement' by an object from the having of one to the having of another property, where those properties belong to the same quality space, and where those properties are such that the object's successive havings of them implies that the object changes non-relationally. (Lombart 1986: 114)[16]

This fits nicely with the proposed mathematical formalisation: The random variables Y_t refer to properties of things. These properties are represented by a definite set $\mathcal{Y}$, the 'property space'. And events are changes in things represented by variables, $\{Y_t - Y_{t-1} \neq 0\}$. Events occur at a definite point in time.[17] Kinds of events may than be described by certain transitions between elements of the set $\mathcal{Y}$, or as classes of such transitions.

But this approach may be at once too general and too specific to serve its purpose as a general guideline in the social sciences. It may be too specific because the translation of 'event' into 'change of property' does

[16] He argues that *all* events should be treated in this way.

[17] Since an event in general takes some time, it seems inappropriate to say that they occur at a point in time. This creates considerable problems for a formalisation of time sequences, especially if it is based on the continuous-time theory of stochastic process. Hamblin argues that "the time-continuum, modelled on the real numbers, is richer than we need for the modelling of empirical reality." (cited in Galton 1984: 19)

not capture the most general idea of event playing the role of an efficient cause. Parties, strikes, wars etc. are certainly events, and they generally are considered to have causal efficiency. Still, the notion of an event as a change of a property can only be adapted to such events at the price of some distortion of the event under study.

On the other hand, the notion of events and their probabilistic interdependence in time does still not capture the realist notion that causes should reflect mechanisms (Sørensen 1998), capacities (Cartwright 1989) or productivity (Bunge 1963). Events as changes of state may not involve mechanisms. They may only be sequences like sunset after sunrise, so that the concept of events as changes of properties of things is too general.

5.7. Agents, Actions, and Events

In the words of Bunge (1963: 46, his italics), "the reduction of causation to regular association, as propounded by Humeans, amounts to mistaking causation for one of its tests .... What we need is a statement expressing the idea—common to both the ordinary and the scientific usage of the word—that causation, far more than a relation, is a category of genetic connection, hence of change, that is a way of *producing* things, new if only in number, out of other things." In statistical discussions, the exhibition of productivity of proposed causes is often side stepped. Instead, many accounts view causality through an analogy with planned, isolated experiments. Experiments are seen as a deliberate manipulation of causes with the goal to provide a magnitude of their effects. This magnitude is perceived as the difference between the value of a measurement on a subject in the presence of the cause and the value of the measurement on the same subject in the absence of the cause. The difference can never be observed and so relates to a counterfactual question. The theory therefore involves constructions of 'similar worlds' to identify such magnitudes.[18]

[18] See Holland (1986), Pratt/Schlaifer (1984), Dempster (1990), Rubin (1990), Galles/Pearl (1998), Pearl (1999) and Robins (1999) for discussions and refinements. These approaches are closer in spirit to J.S. Mill's attempts to codify methods of causal inquiry (Holland 1986: 950) than to elaborations of Humean criteria for the existence of causal links.

Since all the criteria for deriving magnitudes of effects rest on empirically untestable assumptions, they are met with scepticism from statisticians (e.g. Dawid 2000) and sociologists alike. Furthermore, counterfactual accounts are deterministic in that they refer to what would necessarily happen in the presence or absence of the cause. But such a deterministic outlook cannot easily adapt to the variability generally observed in the social sciences.

On the other hand, the insistence on the experimental analogy points to the importance of action based interpretations of causality. In fact, it is sometimes suggested that causality should be defined in terms of human actions and their impact on other humans. Von Wright (1972: Chap. 2.9) distinguishes between doing something and bringing about something and goes on to define *P* as a cause relative to *Q*, and *Q* as an effect relative to *P*, if and only if by doing *P* one could bring about *Q*, or by suppressing *P* one could prevent *Q* from happening. Such a view partly reconciles Bunge's search for productivity with counterfactual analysis: The capacity of a human agent to act and thereby to bring about certain events can hardly be denied. And this capacity includes the possibility of deliberately abstaining from that action. But the power to act and thereby to bring about an event is normally understood to mean that, counterfactually, if the agent would not have acted as, in fact, he did, then the event would not have happened.[19]

Many social phenomena are directly based on actions of individuals or organisations (see e.g. Blossfeld/Prein 1998). As far as sociology is concerned with these phenomena, there is no need to refer to an omnipotent experimenter or to seek rescue in designs that—always imperfectly—mimic the experimental setup of other sciences. Within these fields, sociology does not deal with associations among variables per se, but with events brought about or done by acting people. And claims for causal connection among events brought about by agents can be based on the causal capacities of the agents themselves.

[19] Kelsen (1982) argues that the notion of cause and effect originated from idealised human action and reaction in society, that its origins lie in the projection of crime and punishment, guilt and retaliation, onto nature. An action based reasoning about causal connection would therefore be close to the ancient origins of the idea, but without projecting human capacities on God or nature.

It is tempting to seek the causal connection directly in sequences of actions. But actions should not be treated like events that enter into causal relations as causes and effects. There cannot be a similar connection between actions. Otherwise, as Alvarez/Hyman (1998) point out, one would be led to the idea that agents cause their actions, that actions are events caused by agents. But then "an agent who performs one action performs an infinite series of actions: he causes his action; he causes the causing of his action; he causes the causing of the causing of his action; and so on." (p. 222) We will therefore say that agents cause the result of their action, that they bring about events and that causal connection exist, not between actions, but between events done or brought about by actions.

As Bach (1980) points out, the distinction between actions and events also relieves us from specifying times and places for actions. "Once we have specified all the relevant events in the act sequence and have described them as stemming from a mental episode in the way appropriate to action, we have said all we need to say about which actions were performed and what the agent did." (p. 118)

It is sometimes argued that since human actors act intentionally and behaviour is goal-oriented, the intentions or motives of actors to bring about some effect in the future causes the actor to behave in a specific way in the present. Marini and Singer (1988: 377) say that

> [a] major problem with use of the criterion of temporal order in which behavior occurs, or in which events resulting from behavior occur, is often not a good indication of causal priority. Because human beings can anticipate and plan the future, much human behavior follows from goals, intentions, and motives; i.e., it is teleologically determined. As a result, causal priority is established in the mind in a way that is not reflected in the temporal sequence of behavior or even in the temporal sequence of the formation of behavioral intentions.

But the connection of goals, intentions, and motives to acts and events seems to be much looser than Marini and Singer suggest. In fact, von

Wright (1972: Chap. 3) argues that in the ‘practical syllogism’ of the form: a) person *P* wants to achieve *Y*; b) *P* believes that *Y* can be brought about when he is doing *X*; c) Therefore *P* tries to do *X*; the antecedences a) and b) cannot be understood as causes of the person’s doing *X*. Based on observations of goals, intentions, and motives one may try to give a *teleological* explanation of behaviour. But such explanations are not causal, and they can coexist with causal connections between events that are done and brought about by agents. The fact that social agents can behave intentionally, based on expectations, does not reverse the time order underlying causal statements. The explanandum envisaged by Marini and Singer—why a certain person acts as she chooses to act—is simply different from the explanandum of causal analysis.

5.8. Conceptual Problems

In summary, we propose to investigate causal relations among events employing the concepts of local independence and local autonomy in those cases where one is concerned with events brought about or done by agents. However, the execution of such a program is hampered as well by technical as by philosophical problems. The latter concern the basic concepts of independence and autonomy themselves and we will exhibit some of the more disturbing aspects below.

5.8.1. Local Independence

In his “Foundations of the Theory of Probability” Kolmogorov (1950: 9) says that

> one of the most important problems in the philosophy of the natural sciences is—in addition to the well-known one regarding the essence of the concept of probability itself—to make precise the premises which would make it possible to regard any given real events as independent.

The same applies, we think, to the concept of local independence or similar attempts to provide models for causal independence. Local independence cannot be demonstrated from observations alone.[20] Even though conditional independence as well as local independence are not empirical concepts, it does not follow that they cannot be used in an empiricist program for the assessment of causality. They are needed as regulative ideas in modelling. The role of local and conditional independence is to suggest the kind of relations one needs to take into account, but not to describe the likely results of an investigation.

Second, observed relations between stochastic processes generally depend on the number of processes that are considered. If a further process is included, the local dependence between all processes may change. The theoretical background on which an analysis is grounded will to a certain extent determine the variables and histories to be considered in an analysis. In the words of Suppes

> It is important to emphasise . . . that what is to be taken as background or field will always be relative to the conceptual framework under discussion. In one theoretical approach to the causal analysis of phenomena, the field will include only the consideration of macroscopic bodies and their characteristics, but in another, it will go deeper and consider as well atomic objects and their properties. (1970: 74)

In this sense, there may exist several valid causal analyses based on different sets of stochastic processes. Arguments for the exclusion of certain processes will partly rely on ideas of causal non-dependence, which can be translated into local independence within the mathematical model. On the other hand, the theoretical background will rarely be specific enough to determine exactly what processes are to be considered. In consequence, results cannot be expected to be unique.

Third, and perhaps most disturbingly, the probabilistic concept of local independence does not fully conform with most notions of 'explana-

[20] A similar point has often been made in connection with the role of conditional independence in structural equation models. See e.g. Sobel (1997) and Holland in his discussion of Clogg (1992: 199).

tions'[21]:

- There may be two different histories $H_1 = \{Y_{t-1}, A_{t-1}, \ldots\}$ and $H_2 = \{Y_{t-1}, B_{t-1}, \ldots\}$ that make the respective predictions for Y_t locally independent of $\{X_{t-1}, \ldots, X_0\}$, showing that X_t is at most an indirect cause of Y_t. But neither $H_1 \subseteq H_2$ nor $H_2 \subseteq H_1$ need to hold so that explanations (of spuriousness) are not unique.
- Perhaps even worse, it may happen that Y_t is locally independent of X_t with respect to a history $H_1 = \{Y_{t-1}, A_{t-1}, \ldots\}$, but that it ceases to be so with respect to a larger history, say $H_2 = \{Y_{t-1}, A_{t-1}, B_{t-1} \ldots\}$. Including more information for the prediction of Y_t might destroy local independence. The work of Clogg and Haritou (1997) contains some valuable examples.
- Finally, local dependence is not antisymmetric. Both processes may be locally dependent on each other. In this respect, the concept is weaker than many accounts of causality would require. On the other hand, if X and Y are mutually locally independent, then under slight regularity conditions (including uncorrelated innovations) X and Y are stochastically independent (e.g. Schweder 1979: 404). But stochastic independence is a much stronger concept than causal unrelatedness would seem to require.

5.8.2. Autonomy

The principle of local autonomy was introduced to ensure that the processes under consideration are not just expressions of a sole underlying process, so that it is meaningful to assess the properties of one process without regard to the other. The condition is formulated in terms of the uncorrelatedness of the martingales M_t^X and M_t^Y, expressing the intuitive notion that what happens next to X, say, should not be directly related to what happens to Y at the same time. The condition excludes two stochastic processes that are functionally related. In that case it may well be that the first process is locally independent of the second,

[21] See Dawid 1979a, 1979b, 1980 for some examples concerning the 'unexpected' behaviour of conditional independence relations.

while the second is locally dependent on the first, but it would contradict common sense to say that the first process is a cause of the second.[22]

But the condition fails in deterministic situations: Granger (1969: 430) used two deterministic processes as an example: If $Y_t = bt$ and $X_t = c(t+1)$, then $V_t^Y = b$ (independent of $X_0, \dots, X_{t-1}$). But one can as well write $V_t^Y = (b/c)(X_{t-1} - X_{t-2})$, which is dependent on $X_0, \dots, X_{t-1}$. Certainly one would not like to call X_t a cause for Y_t even though the martingales corresponding to the two processes are trivially uncorrelated.

Second, in the context of counting process models, the assumption of autonomy is often replaced by the assumption of no common jumps of the two processes. This in turn implies that the martingales M_t^Y and M_t^X are uncorrelated and that $M_t^X M_t^Y$ is again a martingale (Fleming/Harrington 1991: 75). The condition of no common jumps is often easy to handle, but it may obscure somewhat the role of the condition. Schweder (1970), who starts with a common (Markov–) process with state space $\mathcal{X} \times \mathcal{Y}$, uses the assumption of no common jumps in X and Y explicitly as a condition for the existence of processes that can formally be partitioned into processes X and Y.

On the other hand, the condition of no common jumps is often used for quite a different purpose: It can justify the construction of partial likelihoods. But the statistical considerations that lead to the use of only $U_t^Y(Y_t \mid Y_{t-1}, X_{t-1}, \ldots, Y_0, X_0; \theta)$ for likelihood construction and statistical inference should be carefully distinguished from considerations of the role of the two processes within a causal connection. If one is only willing to specify $U_t^Y(Y_t \mid Y_{t-1}, X_{t-1}, \ldots, Y_0, X_0; \theta)$ for statistical purposes one cannot analyse or simulate the dynamics of the compound process even if one might be willing to impute values for all X_t. This point was stressed both by Strotz and Wold (1960) and Cox (1992). Solving the estimation problem by concentrating on only a part of the system, even if justified, need not suffice to answer causal questions.

[22] See Aalen (1987: 188) for an example.

5.9. Conclusions

The discussions on causality, whether originating from a statistical perspective or from the methodology of the social sciences, have only rarely reflected the philosophical insight that a causal connection is a relation between events but not between variables, things, or qualities. Similarly, an agent based theory of causes that suggests itself in many parts of sociology was largely ignored in favour of counterfactual or system theoretic accounts. We have argued here that an agent based idea of causal connections between events can be supplemented by dynamic descriptions of series of events. When series of events of different kinds are represented by autonomous stochastic processes, the absence of a causal connection can be explicated by the concept of local independence. These concepts should be useful at least in those areas of empirical social research that are directly concerned with events brought about by agents.

We have not treated here a problem of central importance: the problem of spurious causes and of confounding. While the dynamic characterisation of series of events seems to allow for a better understanding and a more flexible formulation of these rather intuitive concepts (e.g. Parner and Arjas 1999), the variety of background conditions and situations generally encountered in social research may well preclude a comprehensive theoretical treatment of confounding. An examination of earlier attempts of demonstrating non-spuriousness in sociology, similar to Goldenberg's (1998) article, will certainly enrich further theoretical developments.

5.10. Postscriptum

The preceding paper argues for an approach to causality that characterises causal relations as non-necessary relations between events brought about by agents. The non-necessity which has been advocated forcefully by Anscombe (1971) and others, is here embodied in a probabilistic and dynamic relation between events. And the notion of capacity (Cartwright 1989) or productivity (Bunge 1963) is in a social science context situated in the causal capacities of agents.

Although it was thought that this notion of causality, while rather weak, would nevertheless provide a useful conceptual framework for at least some parts of empirical social research, progress in the clarification of the open problems indicated at the end of the paper has been very slow. Most of the progress, however, is intimately connected with the main theme of the present work, namely the understanding and modelling of incomplete data. This is not too surprising given that the main technical ingredient of the approach depends on dynamic and probabilistic descriptions of events that did not yet occur.

But before I discuss some of the newer results it is necessary to give at least a brief account of a very different notion of causality that not only dominates much of the discussion in philosophy, the social sciences and in statistics, but is also strongly connected to the problem of incomplete data, though in quite a different way. The approach is commonly termed the *counterfactual* approach to causality. It aims to make precise formulations like: Suppose both X and Y happen. Then X is a cause for Y if, counterfactually, had X not have happened, then Y would not have happened. One may similarly insert "is the case that" or "obtain" or "occur" for "happen" in the above formulation and treat the X and Y as facts or propositions or states of affairs instead of events. The philosophical discussion of the counterfactual approach has explored these and several other possible causal relata. But within the statistical formulation (which is sometimes called the *potential response* approach) a causal relation is a relation between variables and not one between events or facts or propositions.

The literature on this notion of causality has grown tremendously during the past few years. To provide a reasonable overview of all aspects of the discussions would require not only a departure from the main arguments of the previous paper but also the introduction of further technical machinery. It should suffice here to mention a few books and review articles that take up most of the current topics discussed in conjunction with this notion of causality. The potential response approach was presented by Pearl (2000). His book features several possible approaches within the counterfactual tradition in econometrics and the social sciences. A much more concise discussion is given in his review from 2003 and a recent update is Halpern and Pearl (2005). Berk (2004: Chap. 5) provides a non-technical exposition of the statistical aspects. Cartwright (1999) and Woodward (2003, 2004) discuss some philosophical issues. A particular feature, the causal Markov condition, is discussed in a series of articles by Hausman and Woodward (1999, 2004), Steel (2005a, 2006a,b), and Cartwright (2002, 2006). Dawid (2000, 2004, 2006) criticises many aspects of the potential response approach from a (Bayesian) statistical view, while Freedman's book (2005) assembles many of his previous critical discussions from a frequentist point of view. Kluve (2004) tries to formulate an account that brings the potential response approach closer to the philosophical counterfactual account. Hoover (2001) presents a "structural" account of causality in macroeconomics which relies heavily on variables connected by structural equations while criticising other aspects of the potential response approach. Spirtes (2005) is a follow up. The potential response approach is now predominant in medical statistics and epidemiology as well. Greenland and Brumback (2002), Greenland (2004, 2005b), Phillips and Goodman (2006), Höfler (2005b, 2006), Hernán (2004) all argue for the use of counterfactual arguments and the potential response approach in particular, and compare it with some older methodologies. Lipton and Ødegaard (2005), in contrast, warn against an overemphasis of causal concepts in epidemiology. In the social sciences, the potential response approach has also gained much influence. See e.g. Gangl and DiPrete (2006) for a review.

The form of the counterfactual approach most often invoked in statistical analyses directly posits (statistical or random) variables as causal relata. It therefore allows a direct connexion with classical statistical methodology and in particular to methods developed for incomplete data. The connexion, very briefly, is this: Let X denote a 'cause', i.e. a random variable with values in $\{0, 1\}$. Also, let (Y_0, Y_1) denote a pair of random variables defined on a common probability space with X, say Ω. They will serve to represent 'effects' of the 'cause' X. Thus, there is a function

$$(X, Y_0, Y_1) \colon \Omega \longrightarrow \{0, 1\} \times \mathcal{Y} \times \mathcal{Y}$$

The interpretation is that Y_0 is 'observed' if the 'event' $\{X = 0\}$ 'occurs', while Y_1 is 'observed' if the 'event' $\{X = 1\}$ 'occurs'.[23] Note that the distribution of Y_0 is *not* the conditional distribution of some variable Y given the 'event' $\{X = 0\}$. Y_0 is defined ('exists') whether or not $\{X = 0\}$ 'occurs', and the same applies to Y_1. The pair (Y_0, Y_1) always 'exists' jointly.[24]

Identifying causal relata with (statistical or random) variables and adopting the language of counterfactuals (including closest possible worlds etc.), the statistician can proceed using statistical techniques developed for incomplete data problems. It is stipulated that observations are coarsened versions of the triple (X, Y_0, Y_1) so that 'observations' are given by the coarsened variables

$$(X, Y^*) \colon \Omega \longrightarrow \{0,1\} \times \big((\mathcal{Y} \times \{\mathcal{Y}\}) \cup (\{\mathcal{Y}\} \times \mathcal{Y})\big)$$

I.e., one either observes Y_0 (if $X = 0$) or Y_1 (if $X = 1$) but never both. Now the 'effect' of the 'cause' X (or $X(\omega)$ or $X(\omega, u)$ etc.) is defined to be some function of the tupel (Y_0, Y_1). Hence, the "'fundamental problem of causal inference"' (Holland 1986: 947) arises: If the function depends

[23] The literature on probability theory generally calls the set $\{X = 0\} = \{\omega \in \Omega \mid X(\omega) = 0\}$ an 'event'. But the classical notion of sets is a static notion. As such, it neither captures dynamics nor does it easily lend itself to notions of agency or autonomy. The rather simplistic equivocation of sets with 'events' is used far too often to make the statistical version of the counterfactual approach palatable. But even when texts do not conflate mathematical objects with real events and changes, the obvious challenge for statistical analysis remains. It was the main difficulty that prompted the current article.

[24] An explication often used refers to possible worlds such that if in fact $\{Y_0 = y, X = 0\}$, then some y' is the value of Y_1 in the possible world closest to the present one, i.e. the possible world that is exactly like the present one except that $X = 1$. How 'closest parallel worlds' are to be interpreted, what criteria there are for 'close', and in particular in what sense these worlds 'exist', has always been a subject of intense philosophical discussion. See Grayling (1982: 68–77) for a succinct review. I have to mention here only that 'existence' in the context of possible worlds is no straightforward concept. And the exposition of the concept becomes even more involved when the entities of reference are random variables within a probabilistic model. In fact, most texts in the statistical literature simply bypass these problems altogether. Nevertheless, they sometimes seem to suggest that a rhetoric based on parallel or possible worlds and reasoning about counterfactual outcomes is the only 'scientific' way to answer causal questions (e.g. Phillips and Goodman 2006).

on both Y_0 and Y_1 (as it should as a measure of departure from Y_0 from Y_1), then its value is never 'observed'.

There is, in this setup, only one obvious solution: The function representing 'causal effects' is linear and the CAR condition holds:

$$(Y_0, Y_1) \perp\!\!\!\perp X$$

In that case,

$$\mathbb{E}(Y_1 - Y_0) = \mathbb{E}(Y_1 \,|\, X = 1) - \mathbb{E}(Y_0 \,|\, X = 0)$$

using both linearity of expectations and independence. But the right hand side can be 'estimated' by replacing expectations (means across $\omega \in \Omega$) by means across the 'population' (means across $u \in \mathcal{U}$).[25]

It turns out that the choice of any type of (statistical or random) variable as causal relata creates a major problem with this view. This is particularly visible in some sociological 'applications'. [26] Two recent examples may illustrate the ambiguities thus created. Harding (2003) published a "counterfactual model of neighborhood effects" on dropping out of school and on teenage pregnancy. A second example is the very detailed and much appraised study of Epstein et al. (2005) on Supreme Court decisions during 'crisis'. But neither neighbourhoods nor crises do something either to young women or to Supreme Court judges. It is plain that young women's living conditions do depend on their neighbourhoods

[25] Even if the counterfactual account of causality is accepted, neither the restriction to linear 'effects' nor replacing differences between $Y_1(\omega, u)$ and $Y_0(\omega, u)$ by some sort of mean is necessitated by a counterfactual conception of causality. In an attempt to capture different 'effects' within different subpopulations (or subsets of Ω) the econometric literature introduced 'heterogeneous' effects (Heckman 2001, 2006, 2008, Heckman and Smith 1997, Heckman et al. 2006). The extension is, however, just obfuscating the original problem. Furthermore, see Fragankakis et al. (2007) for a defence of the standard statistical view.

[26] Shafer (1995: 556-558) also warns against mistaking (statistical or random) variables for causes within his predictive probability tree approach. He gives examples where variables, even if taken only as descriptions of causes, fail to be unique or are not even defined within a part of a probability tree. Moreover, random variables and statistical variables are conflated quite differently across the probability tree. Rather surprisingly, the statistical and applied literature does not pay much attention to these technical difficulties.

and the resources available from the neighbourhood. Similarly, Supreme Court judges certainly are well aware of the political situation and will react to it. But neither the neighbourhood nor 'political crisis' would qualify as a cause in the analysis proposed in the present article: That a cause ought to be identified with the event brought about by an act of an actor. Now, even with a liberal interpretation of an 'event brought about by … ' probably neither 'crises' nor 'neighbourhoods' would qualify as causes. There is more to the examples than just a demonstration of a difference in definition (or personal preferences): If 'cause' should refer to more than just some 'living conditions' or 'political climate', i.e. background information; if, in other words, 'causes' are not just an assembly of arbitrary sets of (pre-) conditions but ensembles of states of affairs that necessitate the outcome, then neither a 'political crisis' nor the neighbourhood qualify as causes: Judges do not necessarily judge or vote 'because' of a 'state of emergency' (however defined by Epstein et al., and whether felt or real). And young women act and behave as they do, but not in an explicable way 'because' they grew up in a certain neighbourhood (however 'neighbourhood' is defined by Harding).

There is, in fact, some resistance against the indiscriminant use of any (statistical or random) variable as a 'cause'. Heckman in his review (2006) distances himself from some aspects of the statistical literature on the ground that it relies decisively on the analogy with experimental designs:

> One theme developed in my paper is that major limitations hamper the statistical treatment effect literature in answering important social science questions. These limitations are not surprising since the statistical treatment effect literature is an offshot of the experimental design literature in biostatistics. My essay shows that "technical" assumptions invoked in the statistical treatment effect literature have unappealing implications for social science. (2006: 138)

However, Heckman's review does not depart from the rest of the statistical literature in that the questions and models he proposes refer to counterfactual measures based on the joint distributions of a set of random variables plus the idea that the values of some of the variables can be manipulated. In the contrast to the all embracing approaches of Harding and Epstein et al. (where nearly everything can become a 'cause'), the 'manipulations' envisaged by Heckman are basically the implementations of certain policies. Furthermore, he advocates that the relevance of CAR-type 'assumptions' ought to be judged in the context of (economically or sociologically) reasonable, relevant, and explicit models of self-selection, a requirement that echoes the requirement of Neyman cited in Chapter 3.

Another important difference of this literature to the ideas presented in the present article is that dynamic aspects are central to the article but basically absent in practically all suggestions derived within the counterfactual approach.[27] In contrast, much of the classical philosophical literature always concentrated on events as the relata of causes and therefore relied on dynamic concepts. Even though a few philosophers have taken up the clarifications achieved within the statistical discussion, the impact of philosophers and philosophically informed discourses on the developments within statistics has been minor.[28]

Sociologists often claim another distinctive feature of causes from mere conditions, namely 'mechanisms'. The 'mechanisms' envisaged are however as abstract (or elusive) as the 'selection processes' stressed in the economic literature. To take a recent example, Ní Bhrolcháin and Dyson (2007), discussing the role of causation in demographic research, cite as diverse 'mechanisms' as astrological conjunctions and famines as qualifying as 'mechanisms' 'causing' changes in total period birth rates and other aggregates.[29] The tendency to extend the notion of 'mechanism' beyond its normal use in order to make it a distinguishing feature of causality within the social sciences seems to be at least as far fetched as

[27] Whether authors embrace counterfactual concepts or not, the statistical literature is rather silent with regard to dynamics and models of change. The few recent exceptions are Lok et al. (2004), Bray et al. (2006), Aalen and Frigessi (2007), Neugebauer and van der Laan (2006a,b), Neugebauer et al. (2007), and Moodie et al. (2007).

[28] See in particular Cartwright (2002, 2006), Hall (2004, 2007), Hausman and Woodward (1999, 2004), Hausman (2005), Steel (2005a,b, 2006a,b), and Woodward (2003, 2004) for contributions from philosophers or philosophically inclined statisticians.

[29] On the rather undifferentiated use of the term 'mechanism' and its ramifications see the volume edited by Hedström and Swedberg (1998) and the articles by Sørensen (1998), Machamer et al., Norkus (2005), Opp (2005), Weber (2007) and Steel (2007).

the similar strategy based on self-selection. In any case, the aspirations of most proponents of such a distinction to create a condition as general as possible (sometimes even to comprehensively cover all problems of sociology simultaneously) makes all such efforts rather pointless. In fact, the tendency to embrace nearly everything that can be described at all as a possible mechanism runs counter to any reasonable effort to distinguish coincidences, circumstances, and conditions from actions and events.

A careful review of the recent literature would need to distinguish the 'mechanism' idea from an argument based on 'manipulation', where the latter is mainly based on similarities to 'experiments'. The latter view has been advocated mainly by Pearl (2000, 2003).[30] It is pursued in much of the current literature on 'evidence based medicine', 'effects' of labour market regulations and programs, etc.[31] Since not even the much cited recent works of Heckman helped to draw the distinction between 'experiments' and a reasonable analysis of self-selection processes (conjectured or known, observed or unobserved, random or fixed), it may be permissible to ignore this rather small differences.

However, the approach based on analogues of 'experiments' or 'manipulations' or respective relations between inputs and outcomes mediated by 'mechanisms' might seem to be close to the agent based dynamic concept proposed in this paper. Albeit apparent similarities, the consequences of the different formulations are profound. A manipulation/mechanism etc. might simply involve the daily application of a certain amount of drug *a*. But this may be accomplished in a variety of ways: Allowing a certain group of people to swallow a pill of a certain makeup, injecting a certain amount of the same substance to every one in a group (with or without his or her consent), or designing an 'experiment' that chooses the 'treated' patients by some randomisation device. Advocates of this form of 'experiment' (the 'gold standard' in epidemiology and in studies of 'risks') probably assume that people assigned to either treatment or control will

[30] See also the detailed discussion in Halpern and Pearl (2005).

[31] See Lindley (2002) for critical remarks from a Bayesian point of view and Hacking (1988) for illuminating background information on the origins of statistical 'experiments'.

follow henceforth the prescription given to them without hesitation, or thought, or consideration of intermediate results of the 'experiment', or their own considerations, their personal situation, discussions with relatives and friends, or even perceived well-being. That this is wrong in most circumstances of practical importance just reiterates Heckman's critique in that an evaluation of 'policy interventions' must take into account a reasonable (probabilistic) model of self-selection.

In consequence, it seems unhelpful to insist on the pivotal role of 'controlled experiments' as a a gold standard after 'adjusting' results somehow based on models of self-selection and similar processes. Instead, the whole idea of a fixed, though unknown 'effect' discernable from idealised 'experiments' should to be discredited. [32]

Even though most of the statistical literature focuses on the counterfactual approach to causality, there has been some progress along the lines of problems indicated in the final section of my article. In particular, Didelez (2007, 2008) discusses the definition of local independence and Aalen and Frigessi (2007) make a few helpful comments on the concept of autonomy. Geneletti's contribution (2007) provides a refreshing view of a non-counterfactual view on causality.

[32] The literature on 'causal effects' pursuing the pretended 'gold standard' of randomised experiments is burgeoning. But only a tiny fraction of examples provided by that literature in the social sciences stands up against its own proclaimed aims. Even twin-studies (often invoked to provide a basis for judging 'causality' unambiguously) are probably tainted by self-selection since identical twins are more frequent with older couples (and higher education, and . . .).

6

A Non-parametric Mean Residual Life Estimator. An Example from Market Research

6.1. Summary

The mean residual life (mrl) function dynamically describes the average time to an event, depending on the time since the previous event. It provides a forecast in parallel with the development of the underlying process. From a theoretical point of view, the mrl characterises the distribution of the process completely, but in contrast to other characterisations like the hazard rate, it has a direct interpretation in terms of average behaviour.

We use Kaplan–Meier integrals (weighted averages of residual times) to construct a nonparametric estimator of the mrl. We use results from Stute (1995) and Yang (1994) to describe the asymptotic behaviour of this estimator and derive an approximate variance formula.

We present a small simulation study and apply the estimator (and the variance formula) to data pertaining to purchase time behaviour from the Homescan Panel™, A. C. Nielsen, Germany.

6.2. Introduction

The mean residual life (mrl) function dynamically describes the average time to an event, depending on the time since the previous event. As an important example for functionals of the Kaplan–Meier estimator it has been studied by many authors, e. g., Gill (1983), Gijbels and Veraverbeke (1991), Yang (1994) and Stute (1995). From a theoretical point of view, the mrl characterises the distribution of the process completely, see e.g. Shaked and Shanthikumar (1991). The mrl function can therefore be used in model formulation just as densities or hazard functions are. The mrl function is defined as a conditional expectation of the time to an event given that that time is larger than a given value. Its computation thus involves integrals over unbounded intervals of the real line. While conditional expectations are easily interpreted and often the object of immediate interest in applications, the fact that integrals over unbounded domains are involved severely hampers the analysis of estimators in the presence of censoring. This issue has been discussed extensively by Stute and Wang (1993), where a strong law of large numbers is given for such functionals. Instead of integrating with respect to the cumulative hazard rate – as was done in previous work – the authors use integrals with respect to the distribution function of the underlying random variable. Along the same lines, Stute (1995) provides a central limit theorem in this situation. A crucial role is played by the Kaplan–Meier weights, see Stute and Wang (1993: 1593) and Stute (1995: 423), respectively. These quantities generalise the well-known weights for non-parametric estimation in sampling theory, where the inverse inclusion probabilities are used as weighting factors. In the presence of independent censoring, however, the suitable weights are stochastic and vary between the observations.

6.3. An Example from Market Research

A central goal of market research is to describe the market performance of ”fast moving” consumer goods (fmcg). Those are products which are perishable or quickly used up, like food or detergents in contrast to cars, washing machines etc. Throughout this paper, we primarily have fmcg in mind when we talk about products. Both manufacturers and retailers have a strong interest in identifying the more or less successful items of a product class (pc). One way of doing this is to collect data on the purchase behaviour of households. Specifically, we may observe the purchase acts of the participating consumers for a pc of interest during a fixed period, say one year. We call the duration between two consecutive purchase acts by the same household an interpurchase time (ipt). As a pc consists of several items, different types of ipt occur: On the aggregate level, we have the time between two purchases in the pc, while for each product, there is also an item-specific ipt starting and ending with a purchase of the specific article.

In order to reduce problems of dependencies between observations we use only *one* ipt of each type per household in the sample. When interpreting the data, we also have to be aware of possible censorings. They occur at the end of the observation period, in case no repurchase has taken place by then.

Statistical inference concerning durations falls into the realm of survival analysis, where the most common quantity is the hazard rate. In this paper, however, we will focus on the mean residual life function (mrl) instead: Denote by T the random variable describing the length of the ipt and by F its cumulative distribution function. Then the mrl $m(t)$ at some time t equals the average remaining time to repurchase, given this event has not yet taken place:

$$m(t) \coloneqq \mathbb{E}_T(T - t \mid T > t) = \frac{\int_t^\infty u - t \, dF(u)}{1 - F(t)} \tag{6.1}$$

We will discuss the mrl in more detail in Section 6.4. For the moment, we only note that the mrl is a conditional expectation and thus describes the mean behaviour of consumers. This feature makes it an interesting quantity for traditional marketing, where it is impossible to focus on specific, single consumers and one has to deal with the whole group of households instead. As an illustration, we present data pertaining to purchase time behaviour from the Homescan Panel™ of A. C. Nielsen, Germany. It is based on approximately 5,800 households surveyed in

2001. We focus on two items of an anonymised pc. Central facts of this pc and the two selected items, ’focus’ and ’competitor’, are given in the next table:

	penetration	market share	proportion of repurchasing HH
Class	0.58	-	0.78
Focus	0.03	0.014	0.50
Comp.	0.14	0.077	0.61

In words, 58% of the German households purchased the pc in 2001, 3% bought the focus item and 14% bought the competitor item at least once. With market shares of 1.4% and 7.7%, neither of the two products has an overwhelming influence on the development of the pc. On the aggregate level, censoring occurred in 22% of the observations. On the item level, censoring is rather heavy, 50% and 39%, respectively. In Section 6.5, we present a non-parametric estimator for the mrl in this situation, see (6.6). Applied to our data truncated at $\tau^* = 365$ days[1], the resulting curves are given in Figure 6.1.

On a qualitative level, the estimated mrls suggest the following interpretation: The mrl of the pc serves as a reference. To some extent, it provides information on how the average consumer purchases the typical item of this pc. The mrl of the competitor stays rather close to the pc’s mrl. This implies that its consumers perceive this specific item to ’represent’ the pc: whenever the average consumer repurchases the pc, the customer of the competitor’s item repurchases this specific product. Clearly, this makes the competitor a successful article within the pc.

[1] Truncation to $[0, 365]$ means that we used only repurchase times that happened within a year. We are thus looking at the conditional mrl given that repurchase happened within a year. Purchase data from the panel households is available for the following months as well, if we neglect the slight but inevitable panel mortality. Consequently, one might aim at extending the truncation time τ^*, thus diminishing the loss of information. On the other hand, for a household to contribute to the actually used purchase data, it is required that its overall reporting quality exceeds a certain level over the whole period in question. As a result, extending the relevant observation period leads to a decreased sample size presumably offsetting any benefits.

From the focus item's point of view, the situation is considerably less comfortable: Approximately two months (56 days) after the last purchase act, the average remaining time to repurchase of the focus item is 40 days longer than it is for the pc.

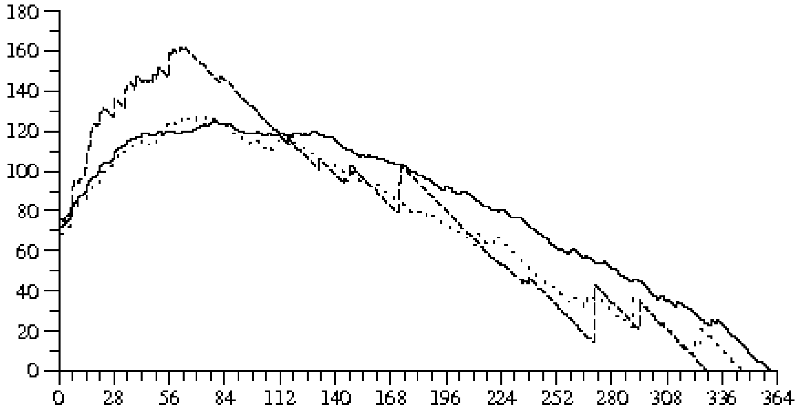


Figure 6.1.: Estimated mrl functions according to (6.6) for the data presented in Section 6.3. Solid: Product class. Dashed: Focus item. Dotted: Competitive item.

To determine whether the differences between the estimated mrl functions really support these arguments or merely occurred 'by chance', we have to introduce a stochastic model of our situation and estimate variances of the mrl estimates as well. We suggest several variance formulae in Section 6.6 and evaluate their performance in section 6.7. Concerning the real life data, it turns out that the corresponding estimations do not render any differences between the mrls significant.

6.4. Mean Residual Life Function

The mrl function defined in (6.1) determines the distribution uniquely. If F is absolutely continuous one can compute the distribution function from the mrl function by

$$1 - F(t) = \frac{m(0)}{m(t)} \exp\left\{ - \int_0^t \frac{1}{m(u)}\, du \right\}, \tag{6.2}$$

see Shaked and Shanthikumar (1991: 614). The mrl function can thus be used as a characterisation of the distribution in the same way as the density or the hazard function can.

As an example, consider a mrl function which is linear on some interval $I = [t_1, t_2]$. For $t \in I$, the mrl thus has the form

$$m(t) = a \cdot (t - t_1) + m(t_1),$$

where $m(t_1) > 0$ and $a \geq -1$. In case $a < 0$, we also have to ensure that $t_2 \leq t_1 - \frac{m(t_1)}{a}$.

For $a \neq 0$, the distribution function on I then equals a Pareto distribution scaled by $F(t_1)$:

$$F(t) = F(t_1) \cdot \left(\frac{a \cdot (t - t_1)}{m(t_1)} + 1 \right)^{-\frac{a+1}{a}} \quad , \quad t \in I$$

Specifically, if $a = -0.5$, the distribution function is linear and the distribution on I is consequently uniform.

For $a = 0$, the distribution function on I is exponential:

$$F(t) = F(t_1) \cdot \exp\left(- \frac{t - t_1}{m(t_1)} \right)$$

Thus, if the mrl function is given on the interval I we get the behaviour of F restricted to I, and complete knowledge of F on I if $F(t_1)$ was known. Note that both formulae simplify considerably in terms of 'local coordinates' $F(t)/F(t_1)$ and $(t - t_1)/m(t_1)$.

In the presence of censoring there may be a point in time after which it is impossible to gather further information. Let τ be the least upper bound on the period of observation. Then the best one can hope for is to estimate integrals up to that point in time. The definition of the mrl function, however, requires an evaluation of the integrals to ∞, not only up to τ. Moreover, the truncation point τ is generally unknown (and not easily estimable).

We will therefore consider to estimate

$$m^*(t,\tau^*) \coloneqq \mathbb{E}(T-t \mid T \in (t,\tau^*)) = \frac{\int_t^{\tau^*} u - t\, dF(u)}{\int_t^{\tau^*} dF(u)} \tag{6.3}$$

for some $\tau^* \leq \tau$. This truncated mrl function m^* need not be related to the mrl function m in any obvious way. However, m^* is still a conditional expectation with an easy interpretation and important applications. Moreover, as a generalisation of (6.2), the truncated mrl function uniquely identifies the distribution function up to τ^* through

$$F(\tau^*) - F(t) = \frac{m^*(0,\tau^*)}{m^*(t,\tau^*)} \exp\left(-\int_0^t \frac{du}{m^*(u,\tau^*)}\right) \tag{6.4}$$

To see this, note that for continuously differentiable (in t) m^*

$$\begin{aligned}
\frac{\partial m^*(t,\tau^*)}{\partial t} &= \frac{\partial}{\partial t}\frac{\int_t^{\tau^*} u - t\, dF(u)}{\int_t^{\tau^*} dF(u)} \\
&= \left(-tf(t) - \int_t^{\tau^*} f(u)\, du + tf(t)\right)\frac{1}{F(\tau^*) - F(t)} \\
&\quad + \frac{f(t)}{(F(\tau^*) - F(t))^2}\int_t^{\tau^*} (u-t)f(u)\, du \\
&= -1 + m^*(t,\tau^*)\frac{f(t)}{F(\tau^*) - F(t)}
\end{aligned}$$

Thus

$$-\frac{\partial}{\partial t}\ln\left(F(\tau^*) - F(t)\right) = \frac{\partial m^*(t,\tau^*)/\partial t + 1}{m^*(t,\tau^*)}$$

so that

$$F(\tau^*) - F(t) = \exp\left(- \int_0^t \frac{\partial m^*(v,\tau^*)/\partial v \mid_{v=u}}{m^*(u,\tau^*)} + \frac{1}{m^*(u,\tau^*)}\, du \right)$$

from which (6.4) follows by noting that the first fraction under the integral is the logarithmic derivative of $m^*(u, \tau^*)$.

6.5. Kaplan–Meier Integrals

With complete data an estimator of the mrl function is

$$\widehat{m}(t) := \frac{\int_t^\infty u - t\, dF_n(u)}{\int_t^\infty \, dF_n(u)} = \frac{\sum_{i=1}^n (t_i - t)\, \mathbb{1}[t_i > t]}{\sum_{i=1}^n \mathbb{1}[t_i > t]} \tag{6.5}$$

With censored observations it is natural to replace F_n by the Kaplan-Meier estimator

$$1 - \hat{F}_n(t) = \prod_{i=1}^n \left(1 - \frac{\delta_{(i)}}{n - i + 1}\right)^{\mathbb{1}[z_{(i)} \leq t]}$$

where $z_i = \min\{t_i, u_i\}$ are the censored observations, $\delta_i = \mathbb{1}[t_i \leq u_i]$ are the censoring indicators and $z_{(1)} \leq \ldots \leq z_{(n)}$ are the ordered values of the observed times while the $\delta_{(i)}$ are the corresponding censoring indicators. For a given upper limit τ^*, the equation (6.5) becomes

$$\widehat{m^*}(t, \tau^*) := \frac{\int_t^{\tau^*} u - t\, d\hat{F}_n(u)}{\int_t^{\tau^*} \, d\hat{F}_n(u)} = \frac{\sum_{i=1}^n (t_i - t)\, \mathbb{1}[t_i > t] w_i}{\sum_{i=1}^n \mathbb{1}[t_i > t] w_i} \tag{6.6}$$

where the w_i are the jump sizes $\hat{F}_n(t_i) - \hat{F}_n(t_i-)$ of the Kaplan–Meier estimator. Note that if the largest observation is censored we will not force $1 - \hat{F}_n$ to be zero after that observation.

In order to analyse the estimator (6.6) we propose a simple stochastic model that will allow the calculation of approximate variances (within that model, of course) and that might be used to gauge the performance of the estimator on real data sets. We will assume that the observations arise from independent and identically distributed copies of the random variables (T, U), where T and U are independent and T has distribution function F while U has distribution function G. Then $1 - H(t) = (1 - F(t))(1 - G(t))$ is the survivor function of the random variable $Z := \min\{T, U\}$. We set $\delta := \mathbb{1}(T \leq U)$ and let $\tau := \inf\{t \mid H(t) = 1\}$ be the least upper bound of the support of Z. In the following we will assume that F is absolutely continuous while we allow for an arbitrary distribution of G. Note, however, that Stute's (1995) results are valid for general F and general G, while Yang's results (1994) allow G to vary with the observations but requires absolute continuous F. Our restriction to absolutely continuous F and identically distributed censoring times is however often applicable and simplifies the formulations considerably. Extending the integrals with respect to $\hat{F}_n$ up to τ requires delicate considerations on the behaviour of $\hat{F}_n$ near τ. The point is well discussed by Stute and Wang (1993) and by Gill (1994). We will here avoid an explicit discussion by either truncating to τ^* or by assuming $\tau = \infty$. In the simulations we will use a distribution function F with a compact support strictly included in the support of G.

6.6. Variance Formulae

The weights in the formula for $\widehat{m^*}$ will in the presence of censoring depend on all the censored observations preceding a given event time. The weights w_i are therefore not independent random variables in our model. There is, however, a general representation of Kaplan–Meier integrals in terms of sums of independent random variables given by Stute (1995: 425). Using that representation, standard results on sums of independent random variables can be used to derive variances of the estimator $\widehat{m^*}$. In the case of absolutely continuous F, the representation simplifies considerably and can be written as

$$\int_0^{\infty} \phi_k(u)\, d\hat{F}_n(u) - \int_0^{\tau} \phi_k(u)\, dF(u) =$$

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\infty \frac{\phi_k(u)-\mathbb{E}(\phi_k(T)\mid T>u)}{1-G(u_-)}\,dM_i(u)+o_P(n^{-1/2}) \tag{6.7}$$

for $k = 1, 2$ and where

$$\phi_1(u) \coloneqq \mathbb{1}[u>t](u-t) \quad , \quad \phi_2(u) \coloneqq \mathbb{1}[u>t]$$

and where

$$M_i(t) \coloneqq \mathbb{1}[t_i \le t, \delta_i = 1] - \int_0^t \frac{\mathbb{1}[z_i \ge u]}{1-F(u)}\,dF(u)$$

is the martingale for the counting process $\mathbb{1}[t_i \le t, \delta_i = 1]$ with respect to the standard filtration (see the Appendix for a derivation).

The representation gives among other results a variance formula via a standard martingale argument (details are deferred to the Appendix):

$$\operatorname{Var}\left(\int_0^\infty \frac{\phi_k(u)-\mathbb{E}(\phi_k(T)\mid T>u)}{1-G(u_-)}\,dM_i(u)\right) =$$

$$\int_0^\infty \frac{\big(\phi_k(u)-\mathbb{E}(\phi_k(T)\mid T>u)\big)^2}{1-G(u_-)}\,dF(u) \eqqcolon \sigma_k^2 \ , \quad k = 1, 2 \tag{6.8}$$

This formula was also derived via a direct argument by Yang (1994).

While the above formulae are useful in theoretical work, practical computations will also have to rely on a direct empirical counterpart of Stute's representation. We start from the ordered values $z_{(i)}, i = 1, \ldots, n$ and the corresponding values $\delta_{(i)}$ and $\phi_{ki} \coloneqq \phi_k(z_{(i)})$. We then define (assuming no ties):

$$\gamma_i \coloneqq \exp\left(\sum_{j=1}^{i-1}\frac{1-\delta_{(j)}}{n-j}\right) \quad , \quad i = 2, \ldots, n, \quad \gamma_1 \coloneqq 1$$

$$a_{ki} \coloneqq \delta_{(i)}\phi_{ki}\gamma_i \ , \quad i = 1, \ldots, n$$

$$b_{ki} \coloneqq \frac{1-\delta_{(i)}}{n-i}\sum_{j=i+1}^{n} a_{kj} \ , \quad i = 1, \ldots, n-1, \quad b_{kn} \coloneqq 0$$

$$c_{ki} := \sum_{j=1}^{i-1} \frac{b_{kj}}{n-j}\ , \quad i = 2, \ldots, n, \quad c_{k1} := 0$$

Finally, we set

$$A_{ki} := a_{ki} + b_{ki} - c_{ki}\ , \quad i = 1, \ldots, n,\ k = 1, 2 \tag{6.9}$$

Writing

$$\widehat{m^*}(t, \tau^*) = \frac{\int_t^{\tau^*} u - t\, d\hat{F}_n(u)}{\int_t^{\tau^*}\ d\hat{F}_n(u)} =: \frac{\hat{A}_1}{\hat{A}_2} \tag{6.10}$$

we get an expression for the variance of the fraction using the delta method:

$$\operatorname{Var}\left(\widehat{m}^*(t, \tau^*)\right) \approx \frac{\hat{A}_{1n}^2 \hat{\sigma}_2^2}{n\hat{A}_{2n}^4} + \frac{\hat{\sigma}_1^2}{n\hat{A}_{2n}^2} - 2\frac{\hat{A}_{1n}}{n\hat{A}_{2n}^3}\operatorname{Cov}(\hat{A}_{1n}, \hat{A}_{2n}) \tag{6.11}$$

where $\hat{\sigma}_1^2$ is the estimated variance of $\hat{A}_1$ and $\hat{\sigma}_2^2$ is the estimated variance of $\hat{A}_2$.

6.7. A Small Simulation Study

The variance approximations of the previous section can be implemented in various ways. In this section we will use simulations to evaluate some of the possible choices. To keep things simple, we use only the exponential distribution with expectation 1 truncated to $[0, 2]$ for F, i.e. $F(t) = (1 - \exp(-t))/(1 - \exp(-2))$ on $t \in [0, 2]$. The truncation assures finiteness of all moments with respect to F and thus all regularity requirements for the validity of the variance formulae are fulfilled. We evaluate the mrl at 0.2, 0.5, 1.0 and 1.5. The mrl at these points is 0.6435, 0.5691, 0.4180 and 0.2293. The survival probabilities $1 - F(t)$ at the evaluation points are 0.79, 0.55, 0.27, and 0.10. As censoring distributions we use the exponential distribution with expectations 5 and 1 corresponding to censoring probabilities of 0.12 and 0.43, respectively.

The corresponding expected proportions at risk are 0.65, 0.33, 0.10, and 0.02 using a censoring distribution with expectation 1, and 0.76, 0.49, 0.22, and 0.08 using a censoring distribution with expectation 5. Sample sizes are 200 and 1000. We use 1000 simulations for each of the combinations.

Table 6.1.: Simulation results for the mrl estimator

Experiment		t=0.2	t=0.5	t=1.0	t=1.5
true mrl		0.6435	0.5692	0.4180	0.2293
n = 200,	Min	0.4637	0.3561	0.1457	0.0020
$U \simeq \exp(1)$	Mean	0.6376	0.5628	0.4087	0.2199
	Max	0.8173	0.7816	0.7404	0.4931
	NA	0	0	0	30
	Var ($\times 10^2$)	0.3130	0.4463	0.6818	0.7556
	as. Var	0.2769	0.3667	0.5123	0.5139
n = 200,	Min	0.5315	0.4427	0.2859	0.0747
$U \simeq \exp(0.2)$	Mean	0.6442	0.5703	0.4178	0.2296
	Max	0.7486	0.7047	0.6058	0.3622
	NA	0	0	0	0
	Var ($\times 10^2$)	0.1534	0.1847	0.2046	0.1588
	as. Var	0.1647	0.1823	0.1886	0.1401
n = 1000,	Min	0.5718	0.4926	0.3319	0.1178
$U \simeq \exp(1)$	Mean	0.6427	0.5688	0.4166	0.2278
	Max	0.7180	0.6598	0.5164	0.3521
	NA	0	0	0	0
	Var ($\times 10^3$)	0.5558	0.7852	1.0237	1.1719
	as. Var	0.5538	0.7334	1.0247	1.0277
n = 1000,	Min	0.5837	0.5144	0.3590	0.1891
$U \simeq \exp(0.2)$	Mean	0.6434	0.5693	0.4177	0.2298
	Max	0.7070	0.6351	0.4843	0.2824
	NA	0	0	0	0
	Var ($\times 10^3$)	0.3353	0.3854	0.3766	0.2688
	as. Var	0.3295	0.3646	0.3773	0.2802

Table 6.1 gives the mrl at the evaluation points together with summary statistics for the estimated mrl from the simulations. The mrl estimator is slightly downward biased for smaller numbers of observations and for

later times. This was to be expected from the supermartingal structure of Kaplan-Meier integrals as discussed by Stute and Wang (1993) and the bias of the Kaplan-Meier estimator itself as given in Fleming/Harrington (1991: 99). Possible consequences and remedies are described by Miller (1983) and Gill (1994: section 8).

In our situation, the asymptotic variances from (6.8) and (6.11) can be computed explicitly. They are also given in Table 6.1. For $n = 1000$, the agreement with the variances estimated in the simulation experiment is excellent. For $n = 200$, however, the variances of the mrl estimates tend to be larger than the asymptotic variances suggest. Note that the variances are multiplied by 100 in the case of $n = 200$ and by 1000 in the case of $n = 1000$.

Turning to the empirical variance estimators, a simple approach would be to use the empirical variance of the terms A_{1i}/B_{2i}, thus avoiding the use of (6.11). Even when using the empirical variances of the A_{ki} separately, one might try simpler versions of (6.11). Yang (1994: 342) proposed to use just the variance of the numerator. Since the denominator is consistent, an appeal to Slutsky's lemma shows this to be a consistent variance estimator. But this would work well only if the expectation of the numerator would be 0. A further possibility is to use both empirical variances, but to ignore the covariance. But the covariance is by definition far from 0. In fact, in our simulation setup, the results using these three estimators had no clear relation with the observed variance of the mrl. We will therefore not present the results in our simulation reports.

A second approach uses the Kaplan–Meier estimator (and its estimated variance) in the denominator of the mrl and in its variance estimator. We will not pursue this possibility here, since we are mainly interested in the performance of Kaplan–Meier integrals. We will therefore also use only the Kaplan–Meier integrals in empirical versions of (6.8).

A third possibility is a direct application of the discrete representation of Stute. This means to compute the terms in (6.9) and use the empirical means, variances, and covariances of these terms in (6.11). We denote this estimator by $\hat{\sigma}_S^2$.

In $\hat{\sigma}_S^2$, one might replace the terms $\hat{A}_{kn}$ by the respective terms used in the mrl estimator. This should give estimators that are better centered and possibly be less variable. We denote this estimator by $\hat{\sigma}_{mrl}^2$.

Lastly, we use an empirical version of Yang's formula (6.8) where we use the estimated mrl for the inner expectation and the empirical Kaplan–Meier integrals for the outer integral. The denominator in the integrand is estimated by the Kaplan–Meier estimator. Moreover, in (6.11), we use the mrl estimators for the $\hat{A}_{kn}$ terms. This version will be denoted by σ_Y^2.

Note that, from a computational point of view, $\hat{\sigma}_Y^2$ requires the computation of the mrl at all (unique) points z_i. In contrast, both $\hat{\sigma}_{mrl}^2$ and $\hat{\sigma}_S^2$ can easily be computed point-wise but are relatively more expensive when the variances are required at all observed points. All three variants, however, are computationally much cheaper than a bootstrap approach.

Table 6.2.: Simulation results for variance estimators, $n = 200$

$U \simeq \exp(1)$		t=0.2	t=0.5	t=1.0	t=1.5
	sim. Var ($\times 10^2$)	0.3130	0.4463	0.6818	0.7556
$\hat{\sigma}_S^2$	Min	0.0012	0.0013	0.0009	0.0000
	Mean ($\times 10^2$)	0.4104	0.5134	0.6052	0.3152
	Max	0.0250	0.0299	0.0313	0.0241
$\hat{\sigma}_{mrl}^2$	Min	0.0012	0.0013	0.0009	0.0000
	Mean ($\times 10^2$)	0.4484	0.5801	0.7441	0.4086
	Max	0.0349	0.0514	0.0532	0.0380
$\hat{\sigma}_Y^2$	Min	0.0011	0.0012	0.0009	0.0000
	Mean ($\times 10^2$)	0.2858	0.3899	0.6235	1.6433
	Max	0.0072	0.0155	0.0512	0.3453

Table 6.3.: Simulation results for variance estimators, $n = 200$

$U = \exp(0.2)$		t=0.2	t=0.5	t=1.0	t=1.5
	sim. Var ($\times 10^2$)	0.1534	0.1847	0.2046	0.1588
$\hat{\sigma}_S^2$	Min	0.0011	0.0012	0.0010	0.0003
	Mean ($\times 10^2$)	0.1675	0.1853	0.1897	0.1438
	Max	0.0041	0.0049	0.0067	0.0060

Table 6.3.: Simulation results for variance estimators, $n = 200$

$\hat{\sigma}^2_{mrl}$	Min	0.0011	0.0012	0.0010	0.0003
	Mean ($\times 10^2$)	0.1682	0.1863	0.1917	0.1474
	Max	0.0044	0.0053	0.0080	0.0097
$\hat{\sigma}^2_Y$	Min	0.0011	0.0012	0.0011	0.0004
	Mean ($\times 10^2$)	0.1686	0.1888	0.2032	0.1959
	Max	0.0023	0.0029	0.0047	0.0101

Looking first at the results for the case $n = 200$ and $U \simeq \exp(0.2)$ (Table 6.2), all three variance estimators are in rather close agreement with the simulated variances, even at $t = 1.5$. Moreover, $\hat{\sigma}^2_S \leq \hat{\sigma}^2_{mrl} \leq \hat{\sigma}^2_Y$ in the mean over all simulations, where also the variability of the estimators increases in this order.

The situation is less favourable in the case of heavy censoring and $n = 200$, given in Table 6.3. Here, all estimators show a large variability. For $t \leq 1$, $\hat{\sigma}^2_Y$ seems to work best. However, at $t = 1.5$, none of the estimators is even close to the variability of the mrl estimator. But note that in this case the asymptotic variance is also not close to the observed variability of the mrl estimator.

Table 6.4.: Simulation results for variance estimators, $n = 1000$

$U \simeq \exp(1)$		t=0.2	t=0.5	t=1.0	t=1.5
	sim. Var ($\times 10^3$)	0.5558	0.7852	1.0237	1.1719
$\hat{\sigma}^2_S$	Min	0.0004	0.0005	0.0006	0.0003
	Mean ($\times 10^3$)	0.6155	0.8133	1.1273	1.1074
	Max	0.0039	0.0052	0.0069	0.0041
$\hat{\sigma}^2_{mrl}$	Min	0.0004	0.0005	0.0006	0.0004
	Mean ($\times 10^3$)	0.6233	0.8281	1.1671	1.1989
	Max	0.0043	0.0061	0.0098	0.0092
$\hat{\sigma}^2_Y$	Min	0.0004	0.0005	0.0006	0.0003
	Mean ($\times 10^3$)	0.5579	0.7436	1.0644	1.2491
	Max	0.0007	0.0010	0.0017	0.0036

Table 6.5.: Simulation results for variance estimators, $n = 1000$

$U \simeq \exp(0.2)$		t=0.2	t=0.5	t=1.0	t=1.5
	sim. Var ($\times 10^3$)	0.3353	0.3854	0.3766	0.2688
$\hat{\sigma}_S^2$	Min	0.0003	0.0003	0.0003	0.0002
	Mean ($\times 10^3$)	0.3302	0.3658	0.3781	0.2826
	Max	0.0004	0.0004	0.0005	0.0005
$\hat{\sigma}_{mrl}^2$	Min	0.0003	0.0003	0.0003	0.0002
	Mean ($\times 10^3$)	0.3305	0.3661	0.3787	0.2834
	Max	0.0004	0.0004	0.0005	0.0005
$\hat{\sigma}_Y^2$	Min	0.0003	0.0003	0.0003	0.0002
	Mean ($\times 10^3$)	0.3311	0.3676	0.3837	0.2978
	Max	0.0004	0.0004	0.0005	0.0005

Turning to the case $n = 1000$ with light censoring (Table 6.4), all three variance estimators are close together and close to the simulated variances. Once again, we find $\hat{\sigma}_S^2 \leq \hat{\sigma}_{mrl}^2 \leq \hat{\sigma}_Y^2$ in the mean over all simulations. We had expected to see $\hat{\sigma}_Y^2$ perform less satisfactorily than the other two estimators since with a small amount of censoring the explicit use of the Kaplan–Meier estimator $1 - \hat{G}$ as a weight in (6.8) might result in unstable behaviour. But this seems not to be the case here.

In the case with heavy censoring ($U \simeq \exp(1)$) we see that $\hat{\sigma}_Y^2$ compares favourably with the other two estimators: It is somewhat closer to the simulated variances and has less variability. To look closer at the problem with the other two estimators, we sampled one of the A_{ki} from each of the 1000 simulation runs. Density estimates of the numerator and denominator variables are given in Figures 6.2 and 6.3. The distributions of the empirical A_k terms are far from normal. They are multimodal with one of the modes close to 0 and they have rather heavy tails. Moreover, the empirical versions of the A_{ki} are dependent so that the variances of the A_{ki} are difficult to estimate accurately. This might explain the larger variability of $\hat{\sigma}_S^2$ and $\hat{\sigma}_{mrl}^2$ compared to $\hat{\sigma}_Y^2$. Moreover, with a larger proportion of censored observations the estimator of the distribution of the censoring variable stabilises and thus also Yang's estimator stabilises.

In conclusion, our limited experience suggests that the variance estimator $\hat{\sigma}_Y^2$ is to be preferred in cases of heavy censoring, while with light censoring all estimators behave similarly.

Lastly, we look at a borderline case where the expectation in (6.8) is finite but where the bias condition (1.6) of Stute (1995: 425) is violated. Suppose F is exponential with expectation 1 and G is exponential with expectation 5. In this case, the mrl of F is constant 1. The function $C(x)$ is given by

$$C(x) = \int_0^{x-} \frac{1}{(1 - H(u))(1 - G(u))}\, dG(u) = \frac{0.2}{1.2}\left(e^{1.2x} - 1\right)$$

and thus

$$\int \phi_k(x)\sqrt{C(x)}\, dF(x)$$

diverges for $k = 1, 2$. With a sample size of $n = 1000$ (Table 6.6), the variance estimators $\hat{\sigma}^2_{mrl}$ and $\hat{\sigma}^2_S$ are rather larger than the simulated variances while $\hat{\sigma}^2_Y$ is still quite close. Moreover, the latter is much less variable than the other two. In fact, looking at the behaviour of the A_{ki}, they show very heavy tails with occasional huge values.

Table 6.6.: Simulation results, $F \simeq \exp(1)$, $G \simeq \exp(0.2)$, $n = 1000$

		t=0.2	t=0.5	t=1.0	t=1.5
mrl	Min	0.8742	0.8424	0.7749	0.7495
	Mean	0.9950	0.9936	0.9906	0.9836
	Max	1.1390	1.1382	1.2160	1.2711
	Var ($\times 10^3$)	1.5778	2.2273	4.0604	7.5099
$\hat{\sigma}^2_S$	Min	0.0010	0.0013	0.0019	0.0025
	Mean ($\times 10^3$)	1.8755	2.7261	5.0650	9.3061
	Max	0.0148	0.0233	0.0551	0.1101
$\hat{\sigma}^2_{mrl}$	Min	0.0010	0.0013	0.0019	0.0025
	Mean ($\times 10^3$)	1.8863	2.7482	5.1377	9.5365
	Max	0.0152	0.0242	0.0587	0.1216
$\hat{\sigma}^2_Y$	Min	0.0009	0.0012	0.0019	0.0025
	Mean ($\times 10^3$)	1.5599	2.2317	4.0548	7.3789
	Max	0.0047	0.0075	0.0196	0.0442

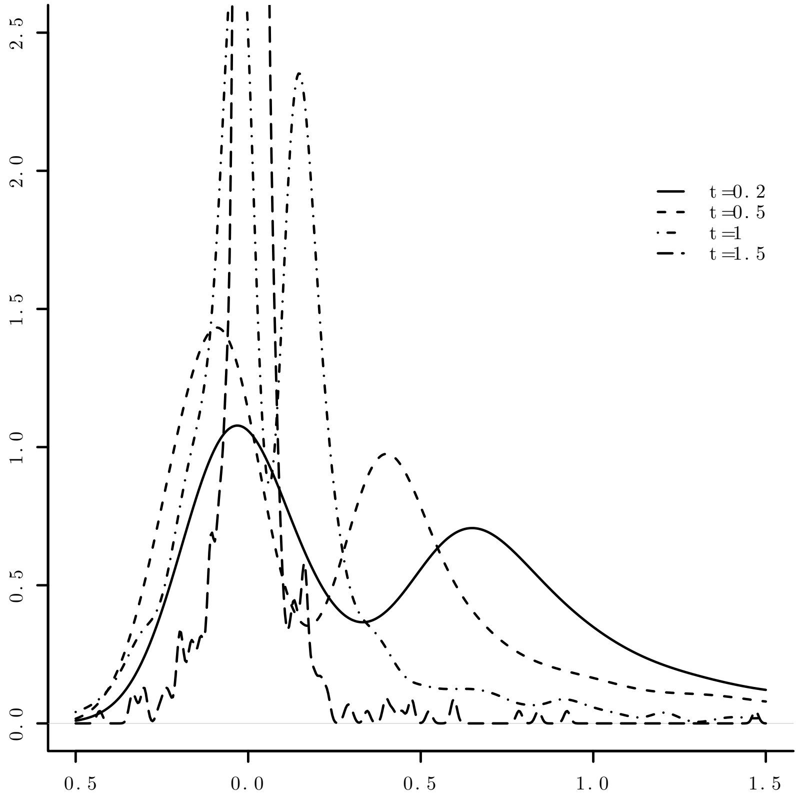


Figure 6.2.: Density estimates for A_{1i}. Solid line: $t = 0.2$, short dashed line: $t = 0.5$, dashed and dotted line: $t = 1.0$, long dashed line: $t = 1.5$.

Table 6.7.: Simulation results, $F \simeq \exp(1)$, $G \simeq \exp(0.2)$, $n = 200$

		t=0.2	t=0.5	t=1.0	t=1.5
mrl	Min	0.7533	0.6608	0.6291	0.4774
	Mean	0.9857	0.9819	0.9772	0.9663
	Max	1.2871	1.3645	1.4910	1.7176
	Var ($\times 10^2$)	0.8320	1.2447	2.1862	4.1630
$\hat{\sigma}_S^2$	Min	0.0032	0.0036	0.0040	0.0043
	Mean ($\times 10^2$)	1.0050	1.4446	2.6152	4.5671
	Max	0.1212	0.1998	0.3797	0.6479

Table 6.7.: Simulation results, $F \simeq \exp(1)$, $G \simeq \exp(0.2)$, $n = 200$

		t=0.2	t=0.5	t=1.0	t=1.5
$\hat{\sigma}^2_{mrl}$	Min	0.0032	0.0036	0.0041	0.0043
	Mean ($\times 10^2$)	1.0286	1.4921	2.7655	5.0157
	Max	0.1368	0.2372	0.5054	0.8394
$\hat{\sigma}^2_Y$	Min	0.0033	0.0037	0.0042	0.0044
	Mean ($\times 10^2$)	0.7962	1.1502	2.1417	4.0573
	Max	0.0429	0.0661	0.1197	0.4599

Looking at a sample size of $n = 200$ (Table 6.7), the mrl has a somewhat larger downward bias. The variance estimators based on the A_{ki} are now rather far from the simulated variances especially at larger t. The estimator $\hat{\sigma}^2_Y$ is closer to the simulated variances. In conclusion, there seems to be some leeway to improve on variance estimators based on the A_{ki}, possibly also in the case of light censoring.

6.8. Appendix

For a function ϕ with $\mathbb{E}(|\phi(T)|) < \infty$ which also meets the appropriately modified moment conditions (1.5) and (1.6) in Stute (1995: 425), Stute (1995) gives a representation of the Kaplan–Meier integral $\int_0^\tau \phi(u)\, d\hat{F}_n(u)$ in terms of sums of independent random variables up to $o_P(n^{-1/2})$. We specialise to absolutely continuous distributions F and G and assume $\tau = \infty$. This rather special case leads to a transparent derivation of the main variance formula and allows to compare the results of Stute (1995) with those of Yang (1994). In particular, Stute's moment condition (1.5) and Yang's condition (ii) (Yang 1994: 339) simply reads

$$\int_0^\infty \frac{\phi(u)^2}{1 - G(u)}\, dF(u) < \infty$$

Stute's representations in the special case can be written

$$\frac{\delta\phi(Z)}{1 - G(Z)} + (1 - \delta)\frac{\mathbb{E}(\phi(T) \mid T > Z)}{1 - G(Z)}$$

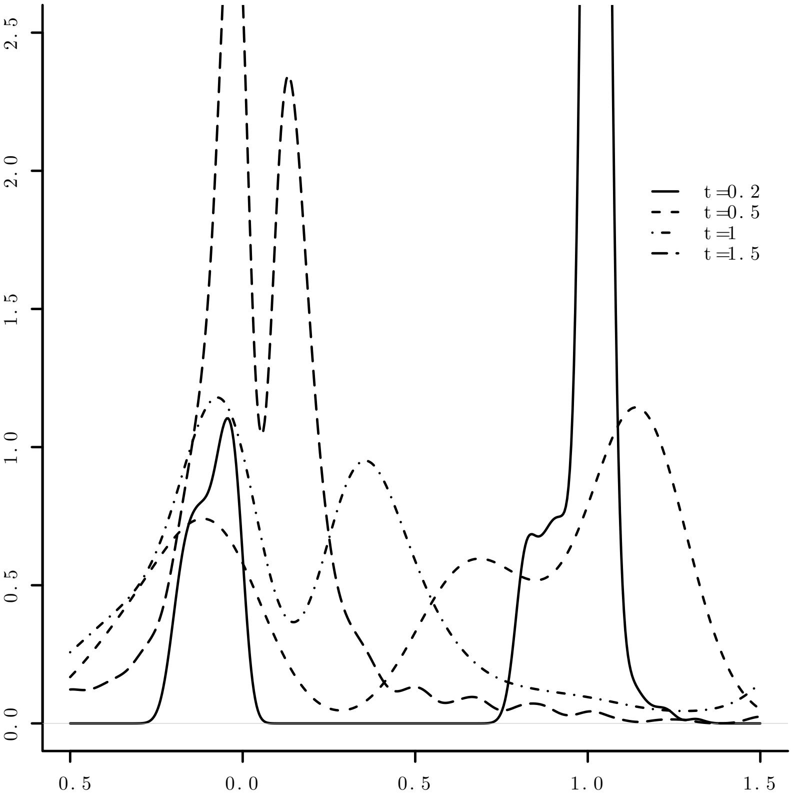


Figure 6.3.: Density estimates for A_{2i}. Solid line: $t = 0.2$, short dashed line: $t = 0.5$, dashed and dotted line: $t = 1.0$, long dashed line: $t = 1.5$.

$$- \iint \frac{\phi(w)\, \mathbb{1}(v < Z)\, \mathbb{1}(v < w)}{1 - H(v)}\, dF(w)\, d\Lambda^G(v) \quad (6.12)$$

where $\Lambda^G(t) := \int_0^t \frac{dG(u)}{1-G(u)}$ is the integrated hazard function of G. The expectation of the above expression with respect to (δ, Z) is easily seen to be $\mathbb{E}(\phi(T))$: Compute the conditional expectation of the first term given $\{T = t\}$ to see that the expectation of that term is $\mathbb{E}(\phi(T))$. For the last two terms, simply write out the expectation with respect to (δ, Z).

Subtracting the expectation $\mathbb{E}(\phi(T))$ and evaluating (6.12) at a fixed argument t we get

$$\begin{aligned}
&\delta\left[\frac{\phi(t)}{1-G(t)} - \frac{\mathbb{E}(\phi(T)\mid T>t)}{1-G(t)}\right] \\
&\quad + \frac{\mathbb{E}(\phi(T)\mid T>t)}{1-G(t)} \\
&\quad - \iint \frac{\phi(w)\,\mathbb{1}(v<t)\,\mathbb{1}(v<w)}{1-H(v)}\,dF(w)\,d\Lambda^G(v) \\
&- \mathbb{E}(\phi(T))
\end{aligned} \tag{6.13}$$

We will start by re-expressing the double integral. For this, note that $\Lambda^H = \Lambda^F + \Lambda^G$ from the definition of the distribution function of Z. Thus,

$$\begin{aligned}
&\iint \frac{\phi(w)\,\mathbb{1}(v<t)\,\mathbb{1}(v<w)}{1-H(v)}\,dF(w)\,d\Lambda^G(v) \\
&\quad = \iint \frac{\phi(w)\,\mathbb{1}(v<t)\,\mathbb{1}(v<w)}{1-H(v)}\,dF(w)\,d(\Lambda^H(v)-\Lambda^F(v)) \\
&\quad = \iint \frac{\phi(w)\,\mathbb{1}(v<t)\,\mathbb{1}(v<w)}{(1-H(v))^2}\,dH(v)\,dF(w) \\
&\qquad - \iint \frac{\phi(w)\,\mathbb{1}(v<t)\,\mathbb{1}(v<w)}{(1-H(v))(1-F(v))}\,dF(v)\,dF(w) \\
&\quad = \int \left(\frac{1}{1-H(\min(t,w))} - 1\right)\phi(w)\,dF(w) \\
&\qquad - \iint \frac{\phi(w)\,\mathbb{1}(v<t)\,\mathbb{1}(v<w)}{(1-H(v))(1-F(v))}\,dF(v)\,dF(w)
\end{aligned}$$

where we used Fubini's theorem in the second equation, and where the first term in the third equation results from a transform of variables.

The last three terms in (6.13) can thus be written as

$$\frac{\mathbb{E}(\phi(T)\mid T>t)}{1-G(t)} - \iint \frac{\phi(w)\,\mathbb{1}(v<t)\,\mathbb{1}(v<w)}{1-H(v)}\,dF(w)\,d\Lambda^G(v)$$

$$
\begin{aligned}
&- \mathbb{E}(\phi(T)) \\
&= \int_t^\infty \frac{\phi(w)}{1 - H(t)}\, dF(w) \\
&- \iint \frac{\phi(w)\, \mathbb{1}(v < t)\, \mathbb{1}(v < w)}{1 - H(v)}\, dF(w)\, d\Lambda^G(v) \\
&\quad - \int_0^\infty \phi(w)\, dF(w) \\
&= \int_t^\infty \frac{\phi(w)}{1 - H(t)}\, dF(w) - \int \left(\frac{1}{1 - H(\min(t, w))} - 1 \right) \phi(w)\, dF(w) \\
&\quad + \iint \frac{\phi(w)\, \mathbb{1}(v < t)\, \mathbb{1}(v < w)}{(1 - H(v))(1 - F(v))}\, dF(v)\, dF(w) - \int_0^\infty \phi(w)\, dF(w) \\
&= - \int_0^t \frac{\phi(w)}{1 - H(w)}\, dF(w) + \int_0^t \frac{\mathbb{E}(\phi(T) \mid T > v)}{1 - H(v)}\, dF(v) \\
&= - \int_0^t \frac{\phi(w)}{1 - G(w)}\, d\Lambda^F(w) + \int_0^t \frac{\mathbb{E}(\phi(T) \mid T > v)}{1 - G(v)}\, d\Lambda^F(v) \\
&= - \int \frac{\phi(w) - \mathbb{E}(\phi(T) \mid T > w)}{1 - G(w)}\, \mathbb{1}(w < t)\, d\Lambda^F(w)
\end{aligned}
$$

Defining

$$
M_i(t) \coloneqq \mathbb{1}[t_i \leq t, \delta = 1] - \int_0^t \frac{\mathbb{1}[z_i \geq u]}{1 - F(u)}\, dF(u)
$$

to be the martingale for the counting process $\mathbb{1}[t_i \leq t, \delta_i = 1]$ with respect to the standard filtration, we can finally write (6.12) for the i–th observation as

$$
\int \frac{\phi(u) - \mathbb{E}(\phi(T) \mid T > u)}{1 - G(u)}\, dM_i(u) \tag{6.14}
$$

In sum, we have the representation

$$
\int_0^\infty \phi(u)\, d\hat{F}_n(u) - \int_0^\infty \phi(u)\, dF(u) =
$$

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\infty \frac{\phi(u)-\mathbb{E}(\phi(T)\mid T>u)}{1-G(u)}\,dM_i(u)+o_P(n^{-1/2}) \tag{6.15}$$

which is just (6.7).

With the last representation at hand it is easy to derive a variance expression using standard martingale arguments:

$$\begin{aligned}
&\operatorname{Var}_{Z,\delta}\left(\int_0^\infty \phi(u)\,d\hat{F}_n(u)-\int_0^\infty \phi(u)\,dF(u)\right)\\
&=\mathbb{E}_{Z,\delta}\left(\left(\int_0^\infty \phi(u)\,d\hat{F}_n(u)-\int_0^\infty \phi(u)\,dF(u)\right)^2\right)\\
&\approx\mathbb{E}_{Z,\delta}\left(\left(\int_0^\infty \frac{\phi(u)-\mathbb{E}(\phi(T)\mid T>u)}{1-G(u)}\,dM(u)\right)^2\right)\\
&=\mathbb{E}_{Z,\delta}\left(\int_0^\infty \left(\frac{\phi(u)-\mathbb{E}(\phi(T)\mid T>u)}{1-G(u)}\right)^2 d\left\langle M,M\right\rangle(u)\right)\\
&=\mathbb{E}_{Z,\delta}\left(\int \mathbb{1}(Z>u)\left(\frac{\phi(u)-\mathbb{E}(\phi(T)\mid T>u)}{1-G(u)}\right)^2 d\Lambda^F(u)\right)\\
&=\iint \mathbb{1}(z>u)\left(\frac{\phi(u)-\mathbb{E}(\phi(T)\mid T>u)}{1-G(u)}\right)^2\frac{1}{1-F(u)}\,dF(u)\,dH(z)\\
&=\int(1-H(u))\frac{\left(\phi(u)-\mathbb{E}(\phi(T)\mid T>u)\right)^2}{(1-H(u))(1-G(u))}\,dF(u)\\
&=\int\frac{(\phi(u)-\mathbb{E}(\phi(T)\mid T>u))^2}{1-G(u)}\,dF(u)
\end{aligned}$$

This is also Yang's (1994) variance formula valid for arbitrary G. We will also need the covariances of the representations for ϕ_1 and ϕ_2. Shortening $\phi_k(u)-\mathbb{E}(\phi_k(T)\mid T>u)$ to $R(\phi_k)(u)$ we have by a similar reasoning as above:

$$
\begin{aligned}
\operatorname{Cov}&\left(\int_0^\infty \frac{R(\phi_1)(u)}{1-G(u)}\,dM(u), \int_0^\infty \frac{R(\phi_2)(u)}{1-G(u)}\,dM(u)\right)\\
&= \mathbb{E}_{Z,\delta}\left(\int_0^\infty \frac{R(\phi_1)(u)R(\phi_2)(u)}{(1-G(u))^2}\,d\left\langle M,M\right\rangle(u)\right)\\
&= \mathbb{E}_{Z,\delta}\left(\int_0^\infty \mathbb{1}(Z>u)\frac{R(\phi_1)(u)R(\phi_2)(u)}{(1-G(u))^2}\,d\Lambda^F(u)\right)\\
&= \iint \mathbb{1}(z>u)\frac{R(\phi_1)(u)R(\phi_2)(u)}{(1-G(u))^2}\,d\Lambda^F(u)\,dH(z)\\
&= \int_0^\infty \frac{R(\phi_1)(u)R(\phi_2)(u)}{1-G(u)}\,dF(u)
\end{aligned}
$$

6.9. Postscriptum

The suggested estimator in this paper is of the form

$$\hat{m}(t) = \int_t^\infty \phi(u)\,dF_n(u)$$

where $F_n(.)$ is the Kaplan-Meier estimator of the distribution function. The analysis then proceeds by replacing the dependent weights of the Kaplan-Meier estimator by a representation using independent contributions. This was first suggested by Yang (1994) and Stute (1995).

These results have recently be extended to cases where in the integral depends on further parameters or when covariates are present. In particular, Sellero et al. (2005) consider estimates of integrals of the form

$$\hat{m}(t) = \int_t^\infty \int \phi(u,v)\,dF_n(u,v)$$

where $F_n(.,.)$ is the joint distribution function to possibly censored variable T and a covariate Y. Uña-Álvarez and Rodrígiuez-Campos (2004) use a smoothed version of the Kaplan-Meier estimator to provide estimators of expectations of bivariate functions of a possibly censored variable and a covariate.

Wang and Jing (2001) and Qin and Zhao (2007) as well as Zhao and Qin (2006) study the application of empirical likelihood methods to Kaplan-Meier integrals.

Delecroix et al. (2008) propose a weighting procedure similar to the ones proposed by Koul et al. (1981) and Leurgans (1987) in order to estimate conditional expectations of censored variables given a covariate.

While there was not much progress in the analysis of strong representations for more general Kaplan-Meier processes there are now some interesting suggestions for estimators of the mean residual life functions and regression versions of it. In particular, Chen and Cheng (2005) and Chen et al. (2005) use inverse probability of censoring weighted extensions of a model first proposed by Maguluri and Zhang (1994). Instead of a proportional mean residual life function, Chen and Cheng (2006) suggested a linear function of the covariates. Jeong et al. (2008) consider estimators of the median residual life based on inverting functions of the Kaplan-Meier estimator.

7

Covariate Effects in Periodic Hazard Rate Models

7.1. Summary

Labour market participation, consumer behaviour, and many other phenomena exhibit strong periodic patterns that result from cyclic behaviour, constraints on the timing of events, or seasonal variation. While these periodicities can generally be neglected when dealing with small data sets or coarsely grouped event times, they pose challenges to the analysis of large data sets with precise recordings. It seems natural to require that statistical models used in the analysis of such data sets reproduce any underlying periodicities. In particular, the conditional hazard rate given covariates should be periodic for all possible values of the covariates. We show that this requirement severely restricts the class of covariate effects models.

We define periodicities by points of zero crossings of the derivative of the hazard rate. We then develop the concepts of hazard envelope and essential extrema. These allow the construction of classes of covariate effect models with time varying coefficients that respect the underlying periodic structure.

7.2. Introduction

Labour market participation, consumer behaviour, and many other phenomena exhibit strong periodic patterns that result from cyclic behaviour, constraints on the timing of events, or seasonal variation. These phenomena become apparent when large data sets with precise recordings of the timing of events become available. Figure 7.1 exhibits the hazard rate of the inter-purchase time of an 1-litre ice-cream package. The estimate is based on data provided by the German Homescan Panel of A.C. Nielsen. The data contain information on the day of purchases for some 8.400 households over a period of three years.

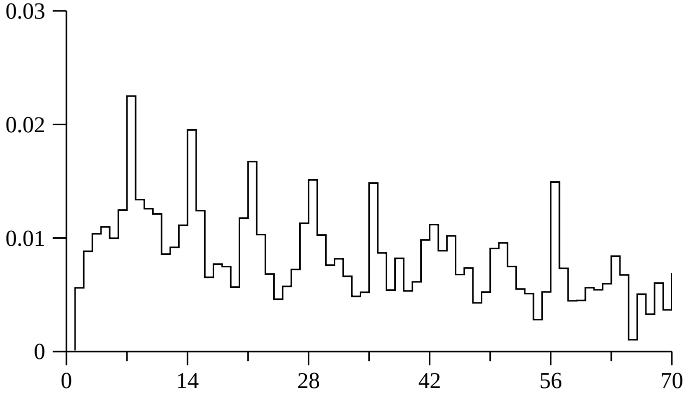


Figure 7.1.: Hazard rate of inter-purchase times (days) of ice-cream packages.

The (discrete) daily hazard rate oscillates with maxima at 7, 14, 21 days and so on. It is plausible to assume that the reason for this behaviour is the weekly purchase schedule of most households. This argument is supported by the fact that these patterns occur across sociodemographic subgroups, e.g. regardless of whether there are children present in the household or not, cf. Figure 7.2.

As a second example, Figure 7.3 presents the estimated hazard rate of

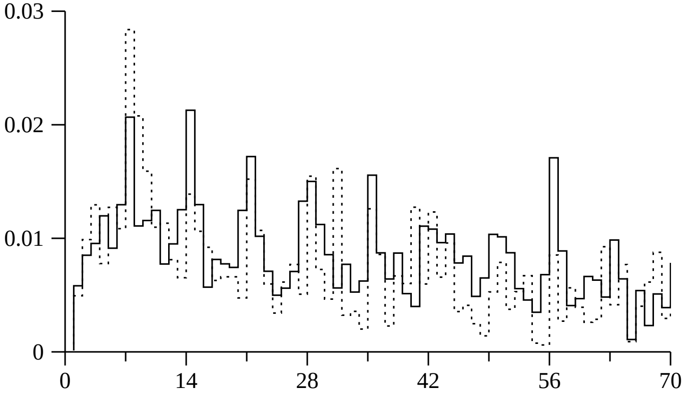


Figure 7.2.: Conditional hazard rate of inter-purchase times (days) of ice-cream packages. Households without children: solid line, Households with children: dotted line.

job durations in Germany for the years 1975 to 1990. The estimate is based on a subsample of records of the social security administration (see Bender et al. 1996) covering some 400,000 job spells. The hazard rate shows large annual peaks, somewhat smaller quarterly peaks and also monthly peaks. However, the number of job durations away from these peaks is still considerable. Again, the findings are similar across subgroups (e.g. men and women). An obvious reason for this pattern is that institutional and juridical regulations in general restrict ending a job to the end of a quarter or to the end of a (calendar) year. Of course, such regulations ought to be the same for all socio-demographic subgroups. In fact, the impact of the regulations is extremely strong: Figure 7.4 depicts the total number of job exits per calendar day for the period 1990–1999. The numbers are based on the complete data of the social security administration. The number of job exits not coincident with a month's end is generally below 100.

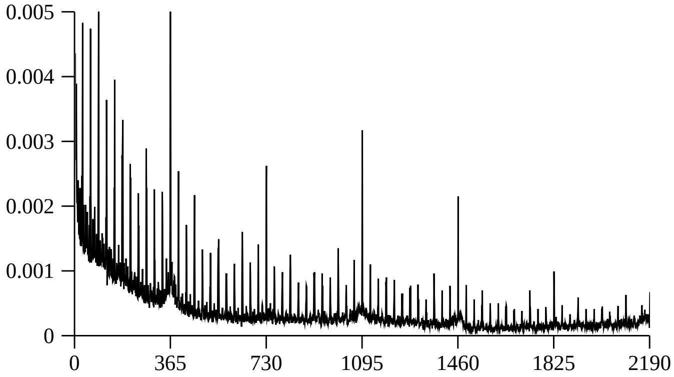


Figure 7.3.: Hazard rate of job durations (days) in Germany 1975–1990.

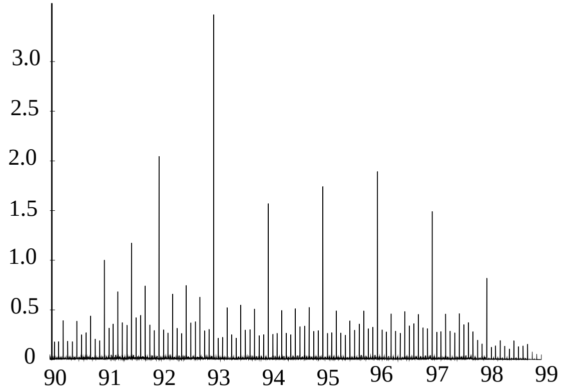


Figure 7.4.: Total number (in million) of job exits by day in Germany 1990–1999. Number for 31.12.1992 truncated.

7.3. Marginal and Conditional Hazard Rates

In both examples, *strong* external influences cause both the conditional and the marginal hazard rates to oscillate. These influences are *common* in the sense that local maxima and minima appear at the same times for the marginal as well as for the conditional hazard rates. Regression models should account for this periodic behaviour: The conditional hazard rates implied by the models should exhibit the same periodicities for all values of the covariates. Moreover, the periodicities of the conditional hazard rates should be the same as those of the implied marginal hazard rates. Surprisingly, however, *no* regression model strictly satisfies these requirements.

Let $T > 0$ and X be random variables on a common probability space representing duration and covariate information. Denote by $\lambda(t) = f(t)/(1 - F(t-))$ and $\lambda(t \mid x) = f(t \mid x)/(1 - F(t - \mid x))$ the marginal and conditional hazard rates. Here, the conditioning is on the events $\{X = x\}$, $f(t)$ and $f(t \mid x)$ are the marginal and conditional densities, and $F(t)$ and $F(t \mid x)$ are the marginal and conditional cumulative distribution functions.

For simplicity, we assume that $\lambda(t \mid x)$ is twice continuously differentiable with respect to t. If $t_1, t_2, \ldots$ are the locations of minima and maxima of $\lambda(t \mid x)$, then $\dot{\lambda}(t_i \mid x) = 0$, $i = 1, 2, \ldots$, where $\dot{\lambda}(t \mid x)$ is the derivative of the conditional hazard rate with respect to t. A possible though rather strict formulation of the above requirements becomes: There is a sequence $0 < t_1 < t_2 \ldots$ such that

$$\dot{\lambda}(t_i) = 0 = \dot{\lambda}(t_i \mid x), \text{ for } i = 1, 2, \ldots \text{ and for all } x. \tag{7.1}$$

To see that in fact no non-trivial model can satisfy this condition, we need to consider the relation between marginal and conditional hazard rates and their derivatives. The marginal hazard rate is given by a time-dependent “convex combination” of the conditional hazard rates:

$$\lambda(t) = \mathbb{E}\big(\lambda(t \mid X) \mid T \geq t\big) \tag{7.2}$$

where the expectation is with respect to the distribution of X conditional on the event $\{T \geq t\}$.

Differentiating this relation leads to

$$\dot{\lambda}(t) = \mathbb{E}\big(\dot{\lambda}(t \mid X) \mid T \geq t\big) + \big[\lambda(t)^2 - \mathbb{E}\big(\lambda(t \mid X)^2 \mid T \geq t\big)\big] \tag{7.3}$$

When the derivatives $\dot{\lambda}(t_i \,|\, x)$ vanish for all x, the first term becomes 0. By Jensen's inequality, the second term is negative unless $\lambda(t \,|\, x)$ is constant in x. Thus, if all derivatives of conditional hazard rates vanish at a point t_i, then the derivative of the marginal hazard rate has to be negative.

To illustrate, consider two subgroups distinguished by the covariate $X \in \{0, 1\}$ and assume

$$\begin{aligned}
\Pr_0 &\coloneqq \Pr(X = 0) = \frac{1}{2} = \Pr(X = 1) \eqqcolon \Pr_1, \\
\lambda_0(t) &\coloneqq \lambda(t \,|\, X = 0) = \frac{5}{4} + \sin t, \\
\lambda_1(t) &\coloneqq \lambda(t \,|\, X = 1) = 2 \cdot \lambda_0(t),
\end{aligned}$$

Then

$$\lambda(t) = \frac{\lambda_0(t) \cdot \exp\big(-\Lambda_0(t)\big) + \lambda_1(t) \cdot \exp\big(-\Lambda_1(t)\big)}{\exp\big(-\Lambda_0(t)\big) + \exp\big(-\Lambda_1(t)\big)},$$

where

$$\Lambda_i(t) \coloneqq \int_0^t \lambda_i(s)\, ds, \;\; i = 0, 1,$$

are the cumulative hazard rates of the subgroups. Figure 7.5 displays the different extrema of the marginal hazard rate $\lambda(t)$ and the conditional hazard rates $\lambda_0(t)$ and $\lambda_1(t)$. Although $\lambda(t) \in [\lambda_0(t), \lambda_1(t)]$ holds due to (7.2), the time-dependence of the convex combination causes the derivative $\dot{\lambda}$ to vanish at different times than $\dot{\lambda}_0$ and $\dot{\lambda}_1$.

7.4. Hazard Envelopes

On the other hand, if the conditional hazard rates $\lambda_i(t), \; i = 0, 1$, oscillate *strongly* and *commonly* in the sense that (w.l.o.g.)

$$\ddot{\lambda}_0(t_{2i}) < 0, \qquad\qquad \ddot{\lambda}_1(t_{2i}) < 0,$$

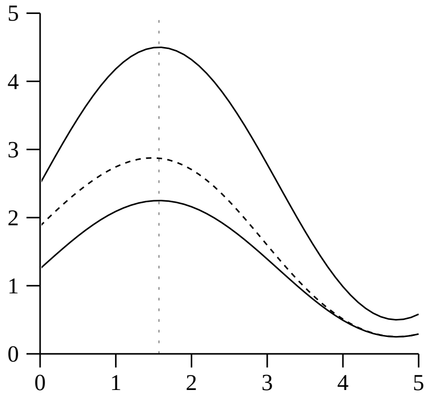


Figure 7.5.: Due to the time-dependence of (7.2), the marginal hazard rate (dashed line) does not have the same extrema as the conditional hazard rates (solid lines).

$$\begin{aligned} \ddot{\lambda}_0(t_{2i-1}) &> 0, & \ddot{\lambda}_1(t_{2i-1}) &> 0, \\ \lambda_0(t_{2i}) &> \lambda_1(t_{2i-1}), & \lambda_0(t_{2i}) &> \lambda_1(t_{2i+1}), \ \forall i \geq 1, \end{aligned}$$

hold, then the marginal hazard rate $\lambda(t)$, being a pointwise convex combination of $\lambda_0(t)$ and $\lambda_1(t)$ as well as a differentiable function of t, must have a local maximum in each interval (t_{2i-1}, t_{2i+1}), $i \geq 1$, and a local minimum in each interval interval (t_{2i}, t_{2i+2}), $i \geq 1$. Qualitatively speaking, $\lambda(t)$ oscillates as well, cf. Figure 7.6.

Thus the marginal hazard rate will oscillate in a similar way as the conditional hazard rates if the latter have common minima and maxima and if they oscillate strongly enough. To bound the behaviour of the marginal hazard rate, we introduce the concept of the *hazard envelope* $(\underline{\lambda}(t), \overline{\lambda}(t))$:

$$\underline{\lambda}(t) \coloneqq \inf_x \{\lambda(t \,|\, x)\} \ , \ \overline{\lambda}(t) \coloneqq \sup_x \{\lambda(t \,|\, x)\} \tag{7.4}$$

Note that in general neither $\underline{\lambda}(t)$ nor $\overline{\lambda}(t)$ need to correspond to any member of the family of conditional hazard rates. Nevertheless, it follows

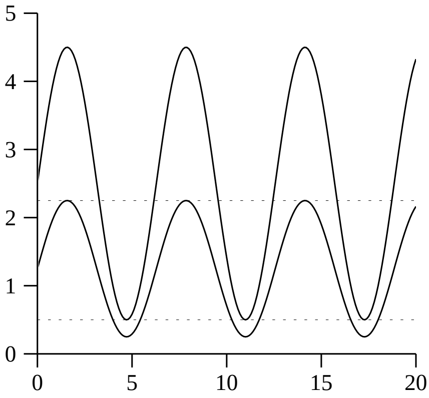


Figure 7.6.: The marginal hazard rate necessarily oscillates between the two dashed lines.

from (7.2) that the marginal hazard rate at a given t is bounded by the extreme points of the conditional hazard rates. Thus

$$\underline{\lambda}(t) \leq \lambda(t) \leq \overline{\lambda}(t) \; \forall t \tag{7.5}$$

Suppose next that maxima of the hazard envelope occur at even numbered times t_{2i} while minima occur at odd numbered times t_{2i-1}. Suppose further that the conditional hazard rates have common minima and maxima at t_{2i} and $t_{2i=1}$, respectively. We say that the conditional hazard rates have an *essential* maximum at t_{2i} if

$$\overline{\lambda}(t_{2i-1}) < \underline{\lambda}(t_{2i}) > \overline{\lambda}(t_{2i+1}) \tag{7.6}$$

We say that the conditional hazard rates have an *essential* minimum at t_{2i+1} if

$$\underline{\lambda}(t_{2i}) > \overline{\lambda}(t_{2i+1}) < \underline{\lambda}(t_{2i+2}) \tag{7.7}$$

An essential maximum implies at least one maximum of the marginal hazard rate in the interval (t_{2i-1}, t_{2i+1}), while an essential minimum implies at least one minimum of the marginal hazard rate in the interval (t_{2i}, t_{2i+2}).

7.5. Consequences for Model Choice

We are now in the position to formulate more reasonable requirements for regression models in situations with strong periodicities: Suppose there is a sequence of time points $t_{2i}, i \geq 1$ at which maxima of the hazards are to occur. Think of the weekly maxima in the hazard rate of inter-purchase times or the quarterly maxima in the hazard rate of job durations. Then one might want to restrict attention to models of covariate effects that, firstly, admit maxima of the conditional hazard rates at the t_{2i} for all values of the covariates, that, secondly, admit the existence of common minima of the conditional hazard rates at some sequence of times t_{2i-1}, and that thirdly, admit essential maxima at all the t_{2i} even in the presence of non-trivial covariate effects.

Consider the class of proportional hazards models with

$$\lambda(t \mid x) = \lambda_0(t)\psi(x\beta), \;\; \psi(x\beta) > 0 \tag{7.8}$$

For this class the envelope hazard coincides with certain conditional hazard rates if the support of the distribution of the covariates is compact. The situation is then very similar to that depicted in Figures 7.5 and 7.6. The first and second requirements are easily met. In fact, they simply depend on the choice of an appropriate baseline hazard rate $\lambda_0(t)$. Whether or not a local maximum is essential will depend both on the extend of covariate effects and the amplitude of the baseline hazard rate. Thus proportional hazard rate models are certainly feasible candidate models.

But what happens if one wants, for some good reason, use non-proportional hazards models? Consider the accelerated failure time model. This model posits a scaling effect of covariates: Suppose that T_x is a random variable representing duration conditional on the covariate value x. Suppose further that there is a random variable T_0 on the same probability space as T_x such that

$$T_x = T_0/\psi(x\beta), \;\; \psi(x\beta) > 0 \tag{7.9}$$

and such that the T_0 have identical distributions for all values of the covariates. The conditional hazard rates are then of the form:

$$\lambda(t \mid x) = \psi(x\beta) \cdot \lambda(\psi(x\beta) \cdot t), \;\; \psi(x\beta) > 0 \tag{7.10}$$

But in such a model, maxima and minima of conditional hazard rates for different values of the covariates cannot coincide, except in the trivial case of no covariate effect, $\psi(x\beta) \equiv 1$.

Does this rule out the use of accelerated failure time models and many other non-proportional hazards models? Not if one is prepared to allow the effects of covariates to change with time. But how does one allow for time-varying covariate effects without destroying the defining features of the accelerated failure time model? After all, if one allows for general time dependent effects $\beta(t)$ and plugs this into (7.10), then the "scaling the time axis" property is destroyed, while there is no obvious way of plugging a time indexed $\beta(t)$ into (7.9).

There is, however, a quite natural way to define changing covariate effects that respects the scaling interpretation of accelerated failure time models. One has to change the global " change of scale" interpretation of covariate effects into a local property at a point in time. That can be done by using derivatives. Starting with the "scale change" interpretations in terms of random variables as in (7.9), one can consider the derivative of the baseline duration with respect to duration with covariate value x. That derivative should be influenced, at a point in time, by the covariate effect at that same point in time. One might thus write

$$\left.\frac{\partial t_0}{\partial t_x}\right|_{t_x=u} = \psi(x\beta(u)) \tag{7.11}$$

But then

$$t_0 = \int_0^{t_x} \psi(x\beta(u))\, du =: \Psi(t_x; \bar{\beta})$$

where $\bar{\beta}$ contains the covariate information and the changes in covariate effects. Therefore

$$T_x = \Psi^{-1}(T_0; \bar{\beta})$$

The hazard rate corresponding to this model of covariate effects is

$$\lambda(t \mid x) = \psi(x\beta(t)) \cdot \lambda_0(\Psi(t; \bar{\beta})) \tag{7.12}$$

Note that this differs from the naive idea to plug in some $\beta(t)$ into (7.10) while it preserves the interpretation of the effects of covariates as a (local) scaling of the time axis.[1]

With this definition of varying covariate effects it is now easy to exhibit versions of the accelerated failure time model that do respect the requirements formulated at the beginning of this section. If we choose

$$\begin{aligned} \Psi(t_{2i}; \bar{\beta}) &= t_{2i} \\ \dot{\Psi}(t; \bar{\beta}) &> 0 \\ \ddot{\Psi}(t_{2i}; \bar{\beta}) &= 0 \text{ and } \dot{\lambda}_0(t_{2i}) = 0 \end{aligned}$$

then

$$\dot{\lambda}(t_{2i} \mid x) = \dot{\psi}(x\beta(t_{2i})) \cdot \lambda_0(\Psi(t_{2i}; \bar{\beta})) + \psi(x\beta(t_{2i}))^2 \cdot \dot{\lambda}_0(\Psi(t_{2i}; \bar{\beta})) = 0$$

With this choice of $\Psi(t)$, $\beta(t)$, and $\lambda_0(t)$, all conditional hazard rates will have the same points of maxima as the baseline $\lambda_0(t)$. As in the case of proportional hazards models, whether the maxima are essential depends on the amplitude of $\lambda_0(t)$ and on the covariate effects process $\psi(x\beta(t))$. But since the envelope hazard will in general not coincide with any of the conditional hazard rates, one needs to compute the envelope hazard explicitly.

7.6. Postscriptum

As far as I am aware of the literature, there is no newer contribution to the analysis of covariate effects in models with periodic behaviour. There have been contributions to the analysis of Poisson point processes

[1] A version of this model for time-dependent covariates has been proposed by Cox and Oakes (1984: 67). Robins and Tsiatis (1992) developed an estimator for this model.

with periodic intensities (see Helmers et al. (2007) for a recent review). However, this literature is silent on the consequences on the form of regression models or the general form of regression models.

8

A Multivariate Buckley-James Estimator

8.1. Summary

Buckley and James (1979) extended the least-squares estimator to cover the case of censored dependent variables. I consider a generalisation of their estimator to the multivariate case based on a non-parametric estimator of the joint distribution of the residuals.

8.2. Introduction

Buckley and James (1979) introduced a regression technique suitable for censored dependent variables. Their estimator uses the least-squares estimating equations and an updating mechanism based on a non-parametric estimator of the residual distribution to deal with the censoring. The procedure is attractive because the use of the least-squares technique allows for an easy interpretation of results and the use of residual analysis, while the updating scheme is general enough to accommodate various forms of censoring and grouping. Consequently, many generalisations of the basic technique have been proposed.

In this paper I explore a possible extension to multivariate dependent variables. Related work, especially that of Lin and Wei (1992), Lee, Wei and Ying (1993), Pan and Kooperberg (1999), and Hornsteiner and collaborators in a series of papers (1996, 1997, 1998), is mainly inspired by the literature on generalised estimating equations. It concentrates on the estimation of the marginal effects of covariates on each of the dependent variables. Accordingly, the least-squares estimating equations are modified to accommodate the multivariate character of the dependent variables. Less emphasis is put on the updating scheme that deals with the censoring problem. The authors suggest to use non-parametric estimators of the marginal distributions of the residuals only. I propose to incorporate the multivariate information from the residual distribution into the updating scheme.

In the next section I introduce Buckley and James' approach to regression estimation with censored observations and, in section 8.4, indicate why it works. Next I consider the multivariate case. Generalisations of the missing information principle are treated in section 8.6. This leads to a multivariate extension of Buckley and James' approach that uses the multivariate information also for the updating scheme. In the final section I examine the performance of the estimator through examples.

8.3. Buckley-James Estimators

Suppose that conditionally on some covariates x, the random variable Y follows a linear regression

$$Y = x\beta + \epsilon \tag{8.1}$$

where x is a $1 \times p$ vector of covariates including a constant, β is a $p \times 1$ vector of unknown regression coefficients, and ϵ is a random variable with mean zero and finite variance. If Y is the logarithm of a positive random variable representing a duration or time to an event, this model is sometimes called accelerated failure time model (Cox/Oakes 1984: Chap. 5.2).

In many applications only censored observations from Y are available. More precisely, suppose that the observations are given by the censored variable Z and censoring indicator δ:

$$Z \coloneqq \min(C, Y)\,,\ \delta \coloneqq \mathbb{1}[C \geq Y]\,,$$

where $\mathbb{1}[.]$ is the indicator function and the censoring variable C is (conditionally) independent of Y. The observations are n independent and identically distributed realizations from (x, ϵ, C). The $n \times (p+2)$ data matrix is given by $(z_i, \delta_i, x_i)_{i=1,\dots,n}$.

In the absence of censoring one can estimate β by minimising the least-squares criterion

$$\sum_{i=1}^{n} (y_i - x_i\beta)^2 = n \int e^2 \, d\hat{F}_n(e) = \sum_{i=1}^{n} \int (y - x_i\beta)^2 \, d\tilde{F}_{ni}(y) \quad (8.2)$$

where $\hat{F}_n(e)$ is the empirical distribution function of the residuals $e_i = y_i - x_i\beta$, and $\tilde{F}_{ni}(y) = \mathbb{1}[y_i < y]$ is the empirical distribution of just one observation y_i.

Miller (1976) and Leurgans (1987), using the second and third representation respectively, proposed replacing the empirical distributions by versions appropriate for censored data. Instead of taking the least-squares criterium (8.2) as their starting point, Buckley and James (1979) suggested to modify the least-squares estimating equations

$$\sum_{i=1}^{n} x_i'(y_i - x_i\hat{\beta}) = 0 \quad \text{or} \quad \sum_{i=1}^{n} x_i' y_i = \left(\sum_{i=1}^{n} x_i' x_i \right) \hat{\beta} \qquad (8.3)$$

In the presence of censoring they proposed to replace the censored observations Z by the conditional expectation of Y given the observed (censored) data Z and the covariates:

$$Y^* = \mathbb{E}_\beta(Y \mid z, \delta, x) = \delta z + (1 - \delta)\mathbb{E}_\beta(Y \mid Y \geq z, x) \qquad (8.4)$$

Note the dependence of the conditional expectation on the unknown parameter β. Replacing Y in expression (8.3) by its conditional expectation gives

$$\frac{1}{n}\sum_{i} x_i' \mathbb{E}_{\hat{\beta}}(Y \mid z_i, \delta_i, x_i) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i' x_i\right)\hat{\beta} \tag{8.5}$$

In other words, the Buckley-James estimator $\hat{\beta}$ solves the normal score function for β when the expectation on the left hand side is computed using $\hat{\beta}$.

Using the model formula (8.1) and a fixed β, an empirical version of the conditional expectation can be evaluated:

$$\begin{aligned}\hat{\mathbb{E}}_\beta(Y \mid z_i, \delta_i, x_i) &=: \hat{y}_i(\beta)\\ &= \delta_i z_i + (1-\delta_i)\hat{\mathbb{E}}_\beta(Y \mid Y_i \geq z_i, x_i)\\ &= \delta_i z_i + (1-\delta_i)\left(x_i\beta + \frac{\int_{e_i}^{\infty} e\, d\hat{F}_\beta(e)}{\hat{S}_\beta(e_i)}\right)\\ &= \delta_i z_i + (1-\delta_i)\left(\sum_{k=i}^{n} v_{ik}(\beta)(z_k - x_k\beta) + x_i\beta\right)\end{aligned} \tag{8.6}$$

where $\hat{F}_\beta$ is an estimator of the distribution function of the residuals (e.g. the Kaplan-Meier estimator), $\hat{S}_\beta$ is the estimated survivor function $1-\hat{F}_\beta$, and I have put

$$v_{ik}(\beta) = \begin{cases} \dfrac{w_k(\beta)}{\hat{S}_\beta(e_i)} & \text{if } e_i < e_k \\ 0 & \text{otherwise} \end{cases}$$

and

$$w_k(\beta) = \hat{P}_\beta(\epsilon = e_k)$$

so that $w_i(\beta)$ is the height of the jump of the estimated distribution at the i-th residual.[1] A solution $\hat{\beta}$ of the estimating equation (8.3) therefore

[1] For ease of notation it is assumed here that the observations are ordered according to the magnitude of the corresponding residuals.

satisfies:

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i' x_i \right)^{-1} \left(\sum_{i=1}^{n} \delta_i x_i' z_i + \sum_{i=1}^{n} (1 - \delta_i) x_i' \hat{y}_i(\hat{\beta}) \right) \tag{8.7}$$

This leads to a straightforward iterative procedure for the computation of $\hat{\beta}$:

1. Assign starting values $\hat{\beta}^0$.
2. Compute $\hat{y}_i(\hat{\beta}^j)$ according to (8.6) using the Kaplan-Meier procedure as an estimator for the distribution of the residuals.
3. Compute $\hat{\beta}^{j+1}$ using the right hand side from (8.7).
4. Go back to step 2 unless some convergence criterion is met.

To be numerically efficient, this simple iterative strategy needs elaboration. Following the steps of the algorithm, the basic choices are:

1. Starting values may be obtained using the least-squares estimator treating all observations as uncensored. This was suggested by Buckley and James (1979). Other choices, e.g. using only uncensored observations, are of course possible, but do not seem to have a decisive influence on the procedure.

2. The Kaplan-Meier estimator is not uniquely defined on the whole real line if the largest residual is censored. Buckley and James suggest to always treat the largest residual as uncensored. This will lead to an underestimation of the regression constant, but should scarcely affect the other regression estimators. Other choices are discussed by Efron (1988), while Lai and Ying (1991) propose to smooth the risk sets.

4. The iteration may not converge to a unique value. This is due to the fact that the right hand side of (8.7) is a piecewise linear function in β. Changing β does not change the weights $v_{ik}(\beta)$ unless the ranks of the residuals change. Therefore, the iterations may oscillate between several values $\hat{\beta}$. The discontinuity of (8.7) hampers the analytic treatment of the estimator. Moreover, the

number of limiting values in finite samples is not predictable, but may potentially be rather large (Currie, 1996). Fortunately, the phenomenon seems to be of practical interest only in rather small samples, in situations where the effect of covariates is small, or when the convergence criterion is very strict (Wu/Zubovic, 1995).[2]

8.4. Score Functions and Censoring

To appreciate why the Buckley-James procedure is a "good" generalisation of estimating equations to censored variables it is helpful to consider it from a more general point of view. Especially the relation between score functions with and without censoring is revealing. Write $\dot{\ell}(\beta) = \dot{\ell}(\beta; Y, x) = x'(Y - x\beta)$ for the score function from the normal linear regression model (8.1). The expectation satisfies

$$\mathbb{E}_\beta(\dot{\ell}(\beta; Y, x)) = 0 \tag{8.8}$$

Moreover, the root $\hat{\beta}$ of the empirical version of the expectation (8.8),

$$\frac{1}{n}\sum_i \dot{\ell}(\hat{\beta}; y_i, x_i) = 0$$

is the maximum likelihood estimator. Even if the distribution is not normal — so that the root of the score function need no longer be a maximum likelihood estimator — $\hat{\beta}$ is consistent and often highly efficient. In the presence of censoring, the censored normal score function $\dot{\ell}^*$ can be expressed as

$$\dot{\ell}^*(\beta; Z, \delta, x) = \mathbb{E}(\dot{\ell}(\beta; Y, x) \mid Z, \delta, x) \tag{8.9}$$

the conditional expectation of the score function with complete observations given the incomplete observations (see e.g. Ibragimov/Has'minskii

[2] Wu and Zubovic (1995) suggested to use the arithmetic mean of all limit values of the algorithm as estimator. This suggestion may be useful in situations where a unique estimator is required (e.g. simulations, using the procedure as building block for more complicated models, etc.). Otherwise, the different values of the limiting cycle of estimators are often very close and it may suffice to report just one of them.

1981: Chap. I.7). This relation between score functions for complete and incomplete observations makes the score function an attractive starting point for the construction of estimators.

It remains to consider the computation of the conditional expectation. From the perspective of the normal linear regression model one might try to use the normal distribution. This was proposed by Schmee and Hahn (1979) and Aitkin (1981). However, one can only expect the good properties of the estimators even outside the normal distribution to extend to censored data situations if the conditional expectation is computed from a non-parametric estimator. In the case of right censored observations, the Kaplan-Meier estimator, being a non-parametric maximum likelihood estimator solving a self-consistency equation, seems to be an appropriate choice. In fact, Lai and Ying (1994), following Ritov (1990) and Severini and Wong (1992), provide a general argument for the use of self-consistent estimators in the computation of conditional expectations for censored and truncated observations.[3] To outline the reasoning it is best to regard the estimation problem as one involving both β and the distribution of ϵ, F, as unknown parameters. Here, β is the parameter of interest and F is treated as a nuisance parameter. In such a context, one may consider the score function corresponding to the profile likelihood. The profile log-likelihood is derived from the log-likelihood $\ell(\beta, F)$ by replacing F with an estimator $\hat{F}_\beta$ treating β as known. It is thus a function of β only. Symbolically, then, one may write

$$\frac{d}{d\beta}\ell(\beta, \hat{F}_\beta) = \frac{\partial}{\partial\beta}\ell(\beta, F)|_{(\beta,\hat{F}_\beta)} + \frac{\partial}{\partial F}\ell(\beta, F)|_{(\beta,\hat{F}_\beta)}\frac{\partial}{\partial\beta}\hat{F}_\beta \qquad (8.10)$$

for its score function. If $\hat{F}_\beta$ is of maximum likelihood type, the sample mean of the second term vanishes. One needs only to consider the score function for β that would result if F was known.

This holds for all unbiased estimating equations for F_β. But an estimator $\hat{F}_\beta$ that maximises the likelihood $\ell(\beta, F)$ in F for β fixed automatically provides an estimator of the least favourable submodel $\beta \mapsto (\beta, \hat{F}_\beta)$ for

[3] The argument extents to estimating equations that are not derived from likelihood functions. See Bickel et al. (1998: Chap. 7.7) for a general discussion of the construction of estimators along these lines.

the estimation of β and therefore (8.10) approximates the efficient score function, making efficient estimation of β feasible. Thus one would like to use an estimator $\hat{F}_\beta$ that simultaneously solves an estimating equation and maximises a non-parametric likelihood.

Taking $F_\beta(u) - \mathbb{1}[Y - x\beta \leq u]$ as a score function for F_β in the uncensored case, one is led via the projection of scores (8.9) to an estimator of F_β that satisfies the corresponding self-consistency equation, namely

$$0 = \mathbb{E}_n \dot{\ell}^*(\hat{F}_\beta) = \mathbb{E}_n(\mathbb{E}_{\hat{F}_\beta \mid Z,\delta,x}(\dot{\ell}(\hat{F}_\beta) \mid Z, \delta, x)) = \hat{F}_\beta(u) - \frac{1}{n}\sum_{i=1}^{n} \hat{F}_\beta(u \mid z_i, \delta_i, x_i) \tag{8.11}$$

But the estimator $\hat{F}_\beta$ that solves the self-consistency equations and maximises the non-parametric likelihood is the Kaplan-Meier estimator. On the other hand, considering $\mathbb{E}(\partial\ell(\beta, \hat{F}_\beta; Y, x)/\partial\beta \mid Z, \delta, x)$ as the profile score function in the presence of censoring, one is led to the estimating equations

$$\begin{aligned}
0 &= \mathbb{E}_n \mathbb{E}_{\hat{F}_{\hat{\beta}}}(\dot{\ell}(\hat{\beta}; Y, x) \mid Z, \delta, x) \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i'(x_i\hat{\beta} - \mathbb{E}_{\hat{F}_{\hat{\beta}}}(Y \mid z_i, \delta_i x_i)) \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i'(x_i\hat{\beta} - \hat{y}_i(\hat{\beta}))
\end{aligned}$$

leading back to (8.5). From this perspective, then, both the choice of the normal score function $\dot{\ell}(\beta) = x'(Y - x\beta)$ as a starting point and the use of the Kaplan-Meier estimator are the appropriate extension of an estimating equation technique to censored data.

8.5. Multivariate Extensions

To consider the multivariate situation I write $\boldsymbol{Y} = (Y_1, \ldots, Y_k)'$ for the column vector of k dependent variables. The covariates are given by a

$k \times kp$ matrix $\boldsymbol{x}$ where the j-th row corresponds to the p covariates x_j of the j-th dependent variable Y_j with zeros padded in the appropriate places. The regression coefficients are given by a column vector $\boldsymbol{\beta}$ of dimension $kp \times 1$. The multivariate linear model can then be presented as

$$\boldsymbol{Y} = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \begin{pmatrix} x_1 & \mathbf{0} & \dots & \dots & \mathbf{0} \\ \mathbf{0} & x_2 & \mathbf{0} & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots & \dots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & x_k \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \boldsymbol{\epsilon} \quad (8.12)$$

with residual vector $\boldsymbol{\epsilon}$. The mean of the residuals is $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ and the covariances are given by $\mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \boldsymbol{\Omega}$. As before, the covariate vectors x_j are assumed to contain a constant. Note that in the case of equal effects $\beta_1 = \beta_2 = \dots = \beta_k$ the $\boldsymbol{x}$ matrix can be reduced to a $k \times p$ matrix.

Now suppose that the data are censored by a k-dimensional variable $\boldsymbol{C} = (C_1, \dots, C_k)'$. Instead of $\boldsymbol{Y}$ only the vectors $\boldsymbol{Z} = (Z_1, \dots, Z_k)' = (\min(Y_1, C_1), \dots, \min(Y_k, C_k))' = \min(\boldsymbol{Y}, \boldsymbol{C})$ and $\boldsymbol{\delta} = (\mathbb{1}[C_1 \geq Y_1], \dots, \mathbb{1}[C_k \geq Y_k])' = \mathbb{1}[\boldsymbol{C} \geq \boldsymbol{Y}]$ are observed. Note that here and in the sequel minima, indicator functions, and (in-)equalities are interpreted component-wise.

To render the conditional distribution of $\boldsymbol{Y}$ identifiable from the censored version $(\boldsymbol{Z}, \boldsymbol{\delta})$ I will assume that the censoring vector $\boldsymbol{C}$ and the vector $\boldsymbol{Y}$ are (conditionally on $\boldsymbol{x}$) independent. Moreover, the support of $\boldsymbol{Y}$ is assumed to be contained in the support of $\boldsymbol{C}$.[4]

Using this model, Lin and Wei (1992), Lee, Wei and Ying (1993), and Hornsteiner and collaborators (1996, 1997, 1998) proposed extensions to the one-dimensional Buckley-James estimator. In these papers, a solution to an equation similar to (8.7) is used. Both Lin and Wei (1992), and Lee, Wei and Ying (1993) use k least-squares estimating equations

[4] Some of the censoring patterns of interest in event history analysis, e.g. censoring of the recurrence times in a semi-Markov process by a fixed observation interval, are not easily represented in this setup. Reference to an underlying process would be necessary to line up censorings and durations according to their timing on a common time scale. See Dabrowska and Lee (1996), Li and Lagakos (1997), and Tsai and Crowley (1998) for some discussion.

disregarding possible correlations. Hornsteiner et al. (1996, 1997, 1998) and Pan and Kooperberg (1999) use a working correlation matrix $V(\alpha)$ (of dimension $k \times k$) in a generalised least-squares estimating equation

$$\sum_{i=1}^{n} \boldsymbol{x}_i' \boldsymbol{V}(\hat{\alpha})^{-1}(\boldsymbol{y}_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}) = \boldsymbol{0} \tag{8.13}$$

in an attempt to gain efficiency. To deal with the censoring, all these proposals use an updating scheme parallel to the one-dimensional case, namely the conditional expectations

$$\begin{aligned} Y_j^{**} &= \mathbb{E}_{\beta_j}(Y_j \mid z_j, \delta_j, x_j) \\ &= \delta_j z_j + (1 - \delta_j)\mathbb{E}_{\beta_j}(Y_j \mid Y_j \geq z_j, x_j), \quad j \in \{1, \ldots, k\} \end{aligned} \tag{8.14}$$

from the j-th model equation. This leads to the correct mean structure while using only the marginal distributions of the residuals. The conditional expectations are then computed from the marginal Kaplan-Meier estimators of the distribution of the residuals. While Lin and Wei and Lee, Wei and Ying simply use the marginal Kaplan-Meier estimators, Hornsteiner (1998) also considers pooled and weighted versions to increase efficiency in certain situations. In addition to the contributions of Y_j^{**} in the updating scheme, both Hornsteiner et al. and Pan and Kooperberg (1999) also base their estimating equations for α on the values of Y_j^{**}. As Hornsteiner (1998: 49) notes, this approach is approximately valid only if the amount of censoring is small.

8.6. The Missing Information Principle and Non-parametric Estimation of Censored Multivariate Observations

Starting with a score function $\dot{\ell}(\boldsymbol{\beta})$ derived from a likelihood $\ell(\boldsymbol{\beta})$, the missing information principle suggests to use the conditional expectation of $\dot{\ell}(\boldsymbol{\beta})$ based on all the available information, not just the information from the marginal distributions. Thus one may consider the conditional expectations

$$\begin{aligned} Y_j^* &= \mathbb{E}_\beta(Y_j \mid \boldsymbol{z}, \boldsymbol{\delta}, \boldsymbol{x}) \\ &= \delta_j z_j + (1 - \delta_j)\mathbb{E}_\beta(Y_j \mid Y_j \geq z_j, (\boldsymbol{z}, \boldsymbol{\delta}, \boldsymbol{x})), \quad j \in \{1, \dots, k\} \end{aligned} \tag{8.15}$$

instead of (8.14). This conditional expectation is based on all the information on $\boldsymbol{Y}$ available from the data while (8.14) uses only the information from the distribution in the j-th dimension. Extending the argument from section 8.4 one would expect (8.15) to give an appropriate generalisation of one-dimensional censored regression if it was possible to exhibit a self-consistent estimator of the multivariate distribution of the censored residuals. Also, from a more practical point of view, it seems advantageous to use as much information as possible in dealing with the censoring process without imposing strong extraneous assumptions. If the degree of censoring is high and if there is considerable correlation within $\boldsymbol{Y}$ or $\boldsymbol{C}$, one might expect (8.15) to perform better than (8.14).

In the context of multivariate proportional hazards models this approach was implicitly suggested by Prentice and Hsu (1997) and Cai and Prentice (1995). On the other hand, this extension has not been discussed in the context of the Buckley-James approach. This is not by accident: in the computation of $\mathbb{E}(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{\delta}, \boldsymbol{x})$ one would need a non-parametric estimator of the joint distribution of $\boldsymbol{\epsilon}$ from censored data that additionally should solve a self-consistency equation, maximise a non-parametric likelihood, and, for practical reasons, should allow for easy computation of conditional expectations along half-lines or orthants.

In dimension 2 or higher, there is no unique self-consistent non-parametric maximum likelihood estimator (NPMLE) of the distribution function of $\boldsymbol{\epsilon}$. In fact, the EM type argument leading to (8.11) will not even result in a consistent estimate. To fix ideas, consider the two-dimensional problem, $k = 2$, disregarding covariates for the moment. Suppose one observes $(z_1, z_2, 0, 1)$, censored in the first component, but exactly observed in the second. This says that the underlying tuple (y_1, y_2) is located on the ray $\{(y_1, y_2) \mid y_1 > z_1, y_2 = z_2\}$ parallel to the first axis. But if the distribution of (Y_1, Y_2) is absolutely continuous, the probability of obtaining another uncensored observation lying on this ray is 0.

Without uncensored observations on the ray there is no empirical support for the computation of the distribution function along this ray. To compute a self-consistent estimator, one needs an expression for $\Pr((Y_1, Y_2) \leq (u_1, u_2) \,|\, (Y_1, Y_2) \in \{(y_1, y_2) \,|\, y_1 > z_1, y_2 = z_2\})$, the last term in the self-consistency equation (8.11) based on a current estimate of the joint distribution. If there are no uncensored observations on the ray, the conditioning event has probability 0 for all sensible starting estimates. Therefore the conditional probability can be defined arbitrarily. But updates of the estimator based on the self-consistency equation will not change due to probability mass transferred from the censored observation to uncensored observations, thus leading to inconsistent estimators.

In response to these difficulties several alternative estimators of the joint distribution of multivariate censored observations have been developed. Pruitt (1993) describes six estimators, summarises their known properties, and compares their small sample behaviour in a limited Monte Carlo experiment. Further comparisons are contained in van der Laan (1997). Some of these estimators are based on a decomposition of the joint distribution into conditional times marginal distributions. The approaches then proceed using the one-dimensional Kaplan-Meier estimator. But the resulting estimators will generally depend on the ordering of the decomposition. Other approaches use smoothing techniques for singly censored observations, thus depending on the choice of a smoothing parameter. The proposals of Dabrowska (1988, 1989) and Prentice and Cai (1992) use special representations of the multivariate survivor function, both representations giving rise to explicit estimators of the distribution function. Gill (1992) provides a lucid introduction to these methods, and both are discussed in Pruitt's 1993 article. Though computationally attractive, both estimators are neither solutions to some self-consistency equation nor are they of maximum likelihood type.

All these approaches may yield negative mass for the increments of the estimated distribution function (Pruitt 1991). This property is especially disturbing when one is interested in computing conditional expectations

$\mathbb{E}_{\hat{F}}(Y_1 \mid Y_1 > z_1, Y_2 = y_2)$ which may result in values $\leq z_1$ for these estimators. Moreover, the implied computation of conditional expectations used in (8.15) are indetermined in general and cannot directly be used in a generalisation of the Buckley-James procedure.

In contrast, there is an essentially unique non-parametric estimator for discrete censored data maximising a likelihood. It was first considered by Campbel (1981a,b). This let van der Laan (1995, 1996, 1997) to consider a non-parametric MLE based on discretised censored observations. In the two-dimensional case, let $D = D_1 \times D_2$ be a rectangle covering the observations so that $(z_{1i}, z_{2i}) \in D$ for all observations. Partition the side of D_1 into q_1 intervals of equal length, $i_{1,l}, l = 1, \ldots, q_1$. Partition D_2 into q_2 intervals $i_{2,l}, l = 1, \ldots, q_2$, also of equal length. This partitions D into $q_1 q_2$ congruent rectangular boxes $i_{1,l} \times i_{2,m}$. Now coarsen the observations as follows: if the observation is uncensored ($(\delta_1, \delta_2) = (1, 1)$) or censored in both dimensions ($(\delta_1, \delta_2) = (0, 0)$), keep the data as $(z_1, z_2, \delta_1, \delta_2)$. If the observation is censored in only one dimension ($(\delta_1, \delta_2) = (0, 1)$ or $(\delta_1, \delta_2) = (1, 0)$), replace the uncensored dimension by the interval it falls into. That is, if the observation is censored in the first dimension, $(z_1, z_2, 0, 1)$, replace z_2 with the interval $i_{2,l}$ to which z_2 belongs. The corresponding (y_1, y_2) are therefore assumed to lie in the strip $\{(y_1, y_2) \mid y_1 > z_1, y_2 \in i_{2l}\}$. Moreover, the strip is restricted to the domain D, $\{(y_1, y_2) \mid y_1 > z_1, y_2 \in i_{2l}\} \cap D$. Similarly, observations only censored in the second dimension, $(z_1, z_2, 1, 0)$, are grouped into $(i_{1,l}, z_2, 1, 0) \cap D$.

In figure 8.1 five observations are depicted. The filled circles represent uncensored observations while the hollow ones represent singly and doubly censored observations. Feasible values of (y_1, y_2) in the case of singly censored observations lie on the rays indicated by solid lines, while values corresponding to the doubly censored observation lie in the orthant indicated by the brocken line. The box around the figure indicates the domain D which is partitioned by intervals of equal length along its two sides. The resulting grid is shown by light lines. The coarsening of the observations does not change the uncensored or doubly censored observations. However, the values of (y_1, y_2) corresponding to the two singly censored observations are now assumed to lie in the shaded strips. While the rays do not contain any uncensored observations, the strip corresponding to the observation censored in the second dimension now contains an uncensored observation. For the reduced data the self-consistency equations contain the term $\Pr((Y_1, Y_2) \leq (u_1, u_2) \,|\, (Y_1, Y_2) \in \{(y_1, y_2) \,|\, y_1 > z_1, y_2 \in i_{2,l(z_2)}\})$ for observations singly censored in the first dimension, where $l(z_2) = \{l \in \{1, \ldots, k\} \,|\, z_2 \in i_{2l}\}$. In general, there will be uncensored observations in the strips corresponding to the conditioning event. Thus, changes in the mass attributed to the uncensored observations will be reflected in the updating scheme for the singly censored observations.

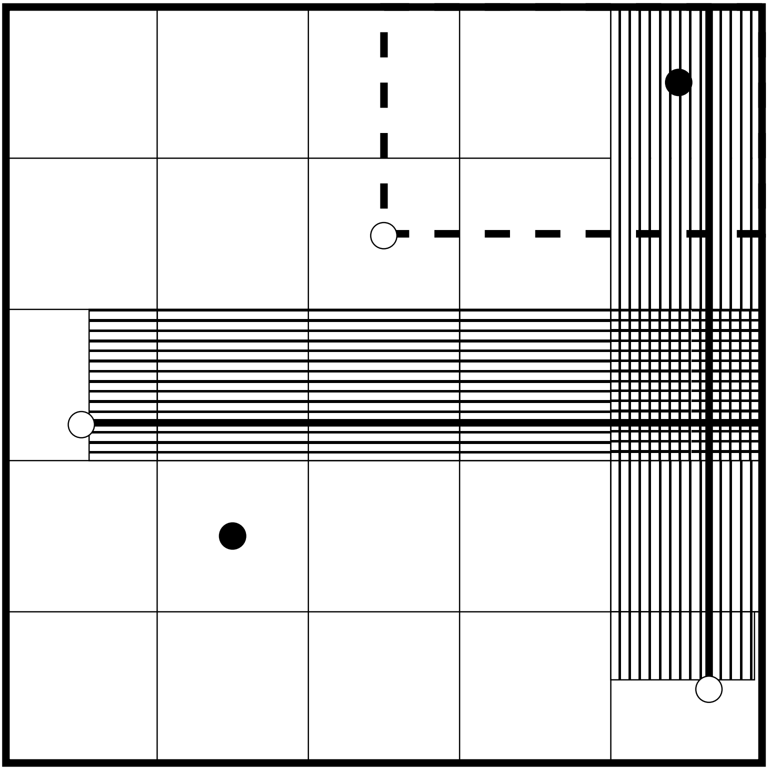


Figure 8.1.: Coarsening censored observations

One may hope that this recaptures the properties of self-consistent estimators in the discrete multivariate and the one-dimensional case, albeit at the cost of throwing away some data.

In fact, van der Laan (1996) showed that the self-consistent MLE based on the reduced data is uniformly consistent and asymptotically normal[5]. To achieve asymptotic efficiency of the reduced data MLE, he shows that the length of the coarsening intervals i in the two-dimensional case have to shrink to 0 at a rate slower than $n^{-1/18}$ (van der Laan 1996: Theorem 5.1). This does not provide much guidance for sample sizes practically encountered. His simulations (1997) suggest that a small interval length of 0.02 for the square $[0,1] \times [0,1]$ and $n = 200$ works well. Our limited experience indicates that in order to attain stable estimates of conditional expectations for the use in Buckley-James iterations it is expedient to use rather larger coarsening intervals.

The procedure is easily generalised to k dimensions. All observations with $0 < \sum_{l=1}^{k} \delta_l < k$ are coarsened to a lattice in $D = D_1 \times \ldots \times D_k$ induced by a partition of the D_j into intervals of equal length. This will ensure that the conditioning events in the self-consistency equations will have positive k-dimensional contents. The estimation procedure for the non-parametric self-consistent MLE of the reduced k-dimensional data can be summarised as follows:

[5] In his simulations and the proofs van der Laan uses a slightly more complicated method of data reduction than the one proposed above. It involves a simultaneous coarsening of the censoring variables $\boldsymbol{C}$ in addition to the coarsening of the uncensored dimensions. If $\boldsymbol{Y}$ is independent of $\boldsymbol{C}$ this is no longer true for the coarsened data version, since $\Pr(Y_1 \in i_{1,l}, \delta_1 = 1) = \Pr(Y_1 \in i_{1,l}, C_1 \geq Y_1) = \int_{i_{1,l}} 1 - G_1(u_-) dF_1(u)$, where F_1 and G_1 are the (marginal) distributions of Y_1 and C_1, respectively. Thus the likelihood no longer factors into a term only containing F and another only depending on the censoring distribution G. Van der Laan's proposal retains the orthogonality between $\boldsymbol{C}$ and $\boldsymbol{Y}$ and thus allows asymptotic arguments based on a sequence of identical models. From a practical point of view and considering that the independence of the censoring scheme cannot be ascertained from the observations one may as well assume that the non-parametric likelihood in the coarsened model factors. One should then bear in mind that different models for the original and coarsened experiment are used, and that one changes models when changing the coarsening grid.

1. Choose a region $D = D_1 \times \ldots \times D_k$. I use $D_l =]\min_i z_{li} - \sigma, \max_i z_{li} + \sigma]$. Note that the choice $\sigma = 0$ will exclude observations that are either right censored in this component at the maximum, or are uncensored in this component at the minimum of the observations.

2. Choose the number q_l of intervals i_l for each dimension l. Partition each side D_l into q_l intervals $i_{l,1}, \ldots, i_{l,q_l}$. I use left open and right closed intervals. Partition D accordingly in $\prod_{l=1}^{k} q_l$ boxes $i_{1,m_1} \times \ldots \times i_{k,m_k}$.

3. Choose starting values. The NPMLE is discrete. It suffices to specify point masses for $\widehat{\Pr}(\boldsymbol{Y} = \boldsymbol{y})$. We choose to put mass $1/n$ on all uncensored observations. The mass of $1/n$ of censored observations is equally spread over the strips implied by the censoring pattern of that observation. To all uncensored observations in the strip and to all intersections of the strip with other strips or with the boundary of D the appropriate part of $1/n$ is added. This will produce a superset of the support points of the NPMLE. Pruitt (1993), Betensky and Finkelstein (1999), and Prentice (1999) discuss the exact determination of the support points of the NPMLE in the two-dimensional case, but the formulation does not easily generalise to higher dimensions.

4. Iterate the self-consistency equations: For each support point $\boldsymbol{y}$ compute the new value $\widehat{\Pr}^{j+1}(\boldsymbol{Y} = \boldsymbol{y})$ as the mean of the conditional probabilities given the observed information, $1/n \sum_i \widehat{\Pr}^{j}(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{Z}_i, \boldsymbol{\delta}_i)$, where the probability of the conditioning event is the sum over the probabilities $\widehat{\Pr}^{j}(\boldsymbol{Y} = \boldsymbol{y})$ lying in the strip determined by $(\boldsymbol{Z}, \boldsymbol{\delta})_i$.

5. Stop the iteration using some convergence criterion. I use the maximum of $|\widehat{\Pr}^{j+1}(\boldsymbol{Y} = \boldsymbol{y}) - \widehat{\Pr}^{j}(\boldsymbol{Y} = \boldsymbol{y})|$ over all support points as convergence criterion.

This EM algorithm generally converges very slowly. Especially the mass of points not in the support of the MLE, but given positive mass by our determination of starting values, decreases only slowly to 0. Prentice

(1999) and Betensky and Finkelstein (1999) proposed to use a direct constraint maximisation algorithm based on the likelihood function. But the approach will fail if the maximum of the likelihood is not unique. This happens if there are strips (or orthants) corresponding to censored observations that intersect D without intersecting other strips or uncensored observations. Region of non-uniqueness can be ascertained in the two-dimensional case, though the procedure is quite tedious. Excluding these region from the maximisation problem would make direct maximisation algorithms very appealing. Unfortunately, we did not find a feasible formulation for the regions of non-uniqueness in the k-dimensional case. In contrast to the direct maximisation approaches the EM algorithm is not hampered by the possible non-uniqueness of the NPMLE. It simply does not change estimates in the regions of non-uniqueness. Since the estimator is to be used repeatedly based on changing data in the Buckley-James procedure, it seems appropriate to use the slow but reliable EM algorithm.

8.7. The Multivariate Buckley-James Estimator

With a NPMLE for the distribution of multivariate censored data at hand, an algorithm for the computation of multivariate regression estimators in the model (8.12) using the Buckley-James approach can be described as follows:

1. Compute starting values for $\boldsymbol{\beta}$. I use the least-squares estimator treating all observations as uncensored.

2. For the $j+1$-th iteration, compute the NPMLE of the residuals based on the data $(\boldsymbol{z}_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}^j, \boldsymbol{\delta}_i)$.

3. Compute new values of the dependent variable as $\boldsymbol{Y}^{*(j+1)} = \hat{\boldsymbol{y}}(\hat{\boldsymbol{\beta}}^j)$ according to (8.15). The conditional expectations of the censored residuals $\boldsymbol{e}_i$ are evaluated as the weighted means of the residuals $\boldsymbol{e}_k$. The estimates from step 2 are used as weights and the summation is over the regions determined by the censoring pattern.

4. Compute new regression coefficients $\hat{\boldsymbol{\beta}}^{j+1}$ using a least squares regression of $\boldsymbol{Y}^{*(j+1)}$ on $\boldsymbol{x}$.

5. Go back to step 2 unless some convergence criterion is met. I use the maximum of $|\hat{\beta}_m^{j+1} - \hat{\beta}_m^{j}| / \max(|\hat{\beta}_m^{j+1}|, 1)$, where m indexes the elements of $\boldsymbol{\beta}$.

The distinctive feature of the estimator is the use of the joint distribution of the residuals to compute expected values in step 3. To illustrate the effectiveness of the computations, I generate $n = 300$ bivariate normal observations with $Y_1 \sim N(0, 1)$, $Y_2 \sim N(0, 1)$ and $\text{corr}(Y_1, Y_2) = 0.8$. These are censored in the second dimension only by $C_2 \sim N(2.4, 1)$. Figure 8.2 compares the estimated expected values of the censored observations (circles) based on the joint distribution (diamonds) with those based on the marginal distribution only (crosses). The estimates based on the joint distribution are clearly better in mimicking the underlying distribution than are the estimates based on the marginal distribution only. Figure 8.3 compares the two approaches in the case of independent components $Y_1 \sim N(0, 1)$, $Y_2 \sim N(0, 1)$, once again with $n = 300$ and $C_2 \sim N(2.4, 1)$. In this situation the estimates based on the joint distribution may be thought to fare less well. While the joint distribution cannot supply any additional information over the marginal distribution, the estimator based on the joint distribution looses information due to the coarsening. In this (and the previous) example I partitioned the first dimension into 10 intervals. It seems apparent from figure 8.3 that the estimates based on the joint distribution do not suffer strongly from the coarsening.

As an example for the effect of joint versus marginal estimation on the regression coefficients I use data from Wei, Lin and Weissfeld (1989: Table 1). The data give natural logarithm of the number of days, z_{li}, to virus positivity in the l-th serum sample of the i-th patient, $l = 1, 2, 3$; $i = 1, \ldots, 36$. There are thus three time dimensions. Patients were treated with ribavirin. There are three treatment groups: placebo, low dose, and high dose. This covariate information is coded in two dummy variables indicating low dose group and high dose group, respectively. There are six observations with missing values in one of the z_{li}. These were excluded from the analysis. Table 8.1 compares the estimated regression

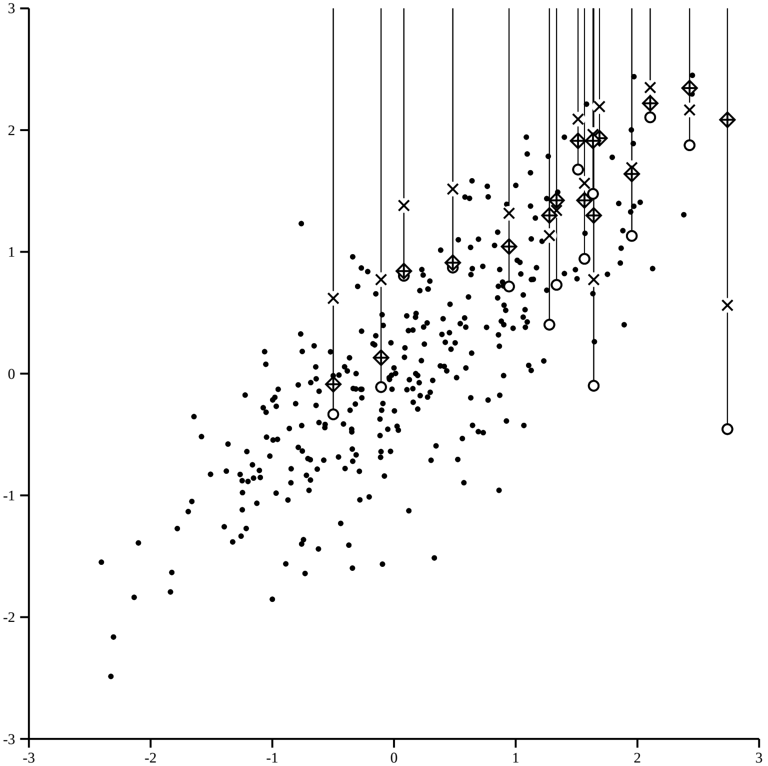


Figure 8.2.: Conditional expectations: Correlation 0.8

coefficients from a model using a marginal Kaplan-Meier estimator with the proposed method using the joint distribution estimator. The latter was computed using a coarsening to five intervals of equal length in each of the three dimensions. The procedure converged after four Buckley-James iterations in each of which the computation of the NPMLE took four to five iterations. The resulting estimated coefficients are all slightly smaller than the coefficients from the marginal estimator.

Table 8.1.: Dependent variable: natural logarithm of days to virus positivity

	marginal	joint

Table 8.1.: Dependent variable: natural logarithm of days to virus positivity

	marginal	joint
Constant 1	1.893	1.893
Constant 2	2.170	2.166
Constant 3	2.179	2.149
low dose 1	0.692	0.674
high dose 1	0.542	0.530
low dose 2	0.168	0.128
high dose 2	0.028	0.021
low dose 3	0.596	0.530
high dose 3	0.252	0.229

8.8. Discussion

The suggested multivariate Buckley-James estimator seems to be a feasible alternative to approaches based on the marginal distribution of the residuals. I have tried it with real and simulated datasets with up to 4000 observations and up to 10 dimensions. The most time consuming part of its computation is the estimation of the joint distribution of the residuals, which may often take 20 to 30 iterations. It would therefore be of interest to develop reliable direct maximisation procedures for the NPMLE.

An obvious obstacle to the use of the estimator is the lack of a variance estimator for regression coefficients. This is due to the fact that there is no variance expression for the NPMLE. Nevertheless, it might be possible to obtain variance estimators from a numerical approximation of the score function.

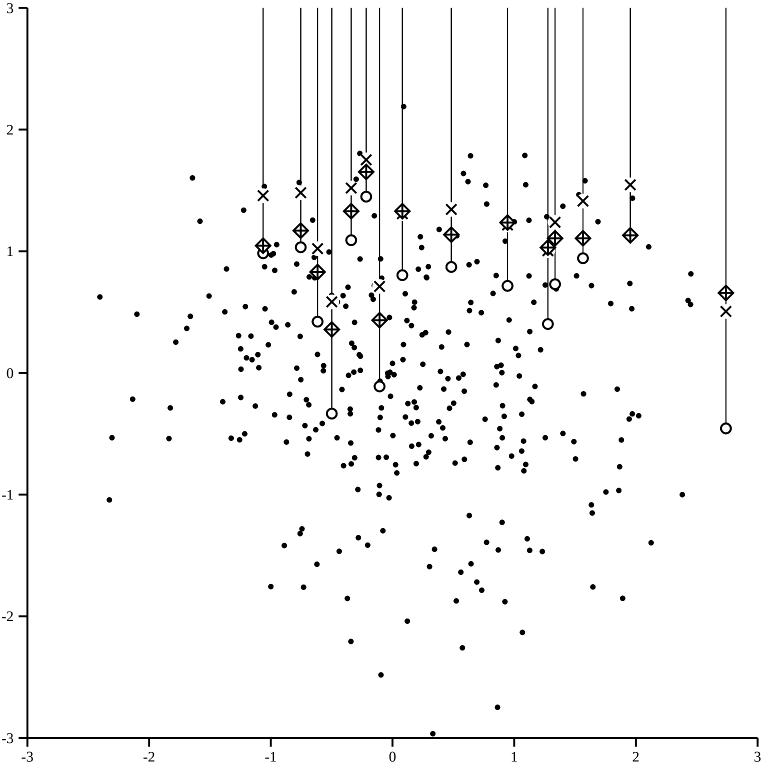


Figure 8.3.: Conditional expectations: Correlation 0

8.9. Postscriptum

The accelerated failure time model and in particular Buckley-James' estimator for one dimensional failure time data has received quite a lot of attention in recent research on regression models with censored data. Johnson (2008) considers variable selection procedures. Kong and Yu (2007) investigate the asymptotic distribution of the Buckley-James estimator under non standard conditions and in particular when the underlying distribution is discontinuous. Other asymptotic results are given by Yu and Wong (2003) and Gørgens (2003) Zhao and Chen (2008) and Zhou and Li (2008) and Zhou (2005) as well as Subramanian (2007) consider empirical likelihood methods. Zeng and Lin (2007) and Jin et al. (2006) as well as Zeng and Lin (2008) explore the construction of efficient estimators. Datta et al. (2007) and Wang et al. (2008) discuss the use of lasso type techniques in the presence of high dimensional convariates. Yu et al. (2007) and Nan et al. (2006) discuss applications under a case cohort design. Lu and Cheng (2007) uses synthetic data approach to construct a partially linear single index model. Huang et al. (2007) and Jin (2007), Fang and Zhao (2006), Ren (2003) as well as Zhou and Wang (2005) discuss robustified versions of the Buckley-James estimator. Leng and Ma (2007) and Heuchenne and van Keilegom (2007a) investigate nonlinear covariate effects in Buckley-James type estimators. General discussions of least squares approaches to censored data problems are provided by Jin et al. (2006) and Heuchenne and van Keilegom (2007b). Rank estimators and smoothed versions of rank estimators are discussed in Chen et al. (2005), Brown and Wang (2007), Peng and Fine (2006), Khan and Tamer (2007), Jin et al. (2003) and Heller (2007). The problem of discontinuity of the objective function of the Buckley-James estimator is investigated by Song et al. (2007). Dimension reduction techniques are introduced by Huang and Harrington (2004, 2005).

8.9.1. Multivariate Approaches

The monograph by Martinussen and Scheike (2006: Chap. 9) discusses recent approaches to the estimation in multivariate censored data. More recent developments include discussions of nonparametric approaches to the estimation of multivariate distributions when data are censored (Aalen et al. 2004, Akritas and Van Keilegom 2003, Van Keilegom 2004, Prentice et al. 2004, Gentleman and Vandal 2002, Alavi and Thavaneswaran 2002, Modarres 2003, Chatterje and Shih 2001, Tien and Sen 2002, Bandeen-Roche and Liang 2002, Wang and Wells 2000, Henriques and Oliveira 2003 and Van der Laan et al. 2002). An emergent field of research is the consideration of semiparametric copula models (Wang 2003, Rivest and Wells 2001, Oakes and Wang 2003, Jiang et al. 2005). Inference for quantiles is discussed by Yin et al. (2003) and Cai and Kim (2003).

Regression models are considered by Van Keilegom and Hettmansperger (2002), Fan and Prentice (2002), He and Lawless (2005), Jin et al. (2006), Lu (2005) and Cai et al. (2008). Oakes and Ritz (2000) discuss the use of copula models in the context of regression and Ivanoff and Merzbach (2004) use the concept of random clouds to discuss aspects of multivariate censored survival analysis. Extensions to more complicated incomplete data models have been discussed by Chen et al. (2008) Antony and Sankaran (2008). Nonparametric kernel regression with censored data was considered by Yu and Lin (2008).

9

On Proportionality of Regression Coefficients in Mis-specified General Linear Regression Models

9.1. Summary

It is shown that estimated regression coefficients in mis-specified general linear regression models (the conditional distribution of the variable of interest given the regressors is a function of a linear combination of the regressors) are approximately proportional to the "true" regression coefficients. The constant of proportionality is computed explicitly and a second order approximation is given. The connexion of this result with similar but stronger findings under the additional assumptions of a convex loss function and normally distributed regressors (Li and Duan 1989: 1009–1052) is clarified.

9.2. Introduction

Over recent years a large amount of empirical evidence—both obtained from real data sets and simulations—has been accumulated for a peculiar stability property of regression models. Using different and possibly mis-specified models results in estimates of regression parameters that are roughly proportional. This is true also if certain kinds of incomplete data are ignored in the analysis. It has long been known that Probit and Logit models often give similar results (Chambers and Cox 1967). D'Agostino et al. (1990) show that a pooled logistic regression model and two variants of Cox' proportional hazards model give nearly identical results for their data sets and offer some reason for the identity. Based on simulations a similar comparison is made by Ingram and Kleinman (1989). Doksum and Gasko (1990) discuss the general relationship between models in binary regression and survival analysis. Addison and Portugal (1987) compare several parametric regressions for a data set on the duration of unemployment and find proportional regression coefficients in all models. Bergström and Edin (1992) give similar results also, using data on unemployment duration. In addition they show that grouping of the dependent variable will attenuate the regression coefficients by a constant factor. Petersen (1991) and Petersen and Koput (1992) treat the effects of grouping in censored regression.

In contrast to the abundance and generality of the empirical findings, theoretical explanations are rare and cover only special cases. Attenuation effects were shown by Bretagnolle and Huber-Carol (1985) for Cox' proportional hazards model if covariates are left out. Ruud (1986) reviews earlier results on proportionality of regression parameters in "limited dependent variable" models with normal regressors. Results for mis-specified correlated Logits were obtained by Neuhaus et al. (1992). Consistency results for linear location-scale models were given by Gould and Lawless (1988). In the case of least squares estimates in a grouped normal regression Stewart (1983) shows that proportionality is asymptotically exact for jointly normal regressors. Chung and Goldberger (1984) generalise his findings to arbitrary information loss under a linearity condition on the reverse regression. More recently, Goldenshluger and Polyak (1993) showed consistency of the least squares estimator in linear models posing very weak conditions on the error term (the models allow for auto-regression and other forms of dependence as well as nearly arbitrary distributions) when the regressors are i.i.d. with a symmetric joint distribution, with expectation 0 and finite fourth moment.

However, there are two general theoretical approaches towards an explanation that seem to cover most of the empirical findings. The first approach places restrictions on the joint distribution of the covariates. It originated from observations on the behaviour of ordinary least squares estimators in nonlinear situations when the regressors are normal (Brillinger 1983). These results were later generalised for models of binary outcomes and other "limited dependent variable" models (see Ruud 1986). Li and Duan (1989) offered the most general proof to date for the class of general linear regression models. These include models for binary outcomes, transformation models, censored regression and generalised linear models.

The second approach is based on a Taylor approximation argument that requires no conditions on the distribution of the regressors. Solomon (1984, 1986) was apparently the first to use it for the comparison of a proportional hazards model when an accelerated life model is true and for an accelerated life model when a proportional hazards model is true. Struthers and Kalbfleisch (1986) gave more general conditions for the behaviour of the partial likelihood estimator (similar to the ones given later by Hjort (1992)), but also used a Taylor approximation argument to get the proportionality result in the partial likelihood case. In a discussion of a paper by Cox and Reid (1987), Skinner (1987) hinted at the possibility of extending the argument of Solomon to the general linear regression case. Later he expanded on the interpretability of proportional regression coefficients (Skinner 1989). In that paper he used the same arguments as Li and Duan for the general linear regression case but did not give a proof via the alternative Taylor approximation.

In this paper I will make Skinner's remarks explicit. The constant of proportionality will be computed and the second order approximation will be given. Then I try to elucidate the relation between Skinner's and Li and Duan's result.

9.3. Notation

The conditional distribution of an outcome variable Y given the values of covariates x describes a *general* linear regression model if the distribution of Y depends on x only through a linear combination of the x. The conditional density (or probability) of Y may be written as

$$f(y \mid \alpha_0 + x\alpha) \tag{9.1}$$

where $f(. \mid .)$ belongs to some class $\mathcal{F}$ of regular conditional densities, x is a $(1 \times p)$ row vector, α_0 is a scalar, and α is the $(p \times 1)$ vector of regression coefficients. It will be assumed that the parameter α is identified at least up to a scalar multiple:

$$\begin{aligned} f(y \mid \alpha_0 + x\alpha) = g(y \mid \alpha_0^* + x\alpha^*) \quad \forall x, y; f, g \in \mathcal{F} \\ \Longrightarrow \quad \alpha^* = \gamma\alpha \,, \; \gamma \in \mathbb{R}\backslash\{0\} \end{aligned} \tag{9.2}$$

The class of general regression models retains the interpretability of relative magnitudes of regression coefficients, a point stressed by Skinner (1989): Let e_i be the i-th unit vector, δ a real scalar. Then

$$\begin{aligned} f(y \mid (x + \delta e_i)\alpha) &= f(y \mid x\alpha + \delta\alpha_i) \\ &= f\left(y \mid x\alpha + \delta\frac{\alpha_i}{\alpha_j}\alpha_j\right) \\ &= f\left(y \mid (x + \delta\frac{\alpha_i}{\alpha_j}e_j)\alpha\right) \end{aligned}$$

so that α_i/α_j, $\alpha_j \neq 0$ is the amount of change in x_j required to achieve an effect on the density equivalent to a unit change in x_i. By the identifiability assumption (9.2) these "equivalent effects" ratios (Skinner 1989) do not depend on the chosen representation, f.

It is sometimes useful to describe the model given by (9.1) in terms of random variables instead of densities. Let us write

$$Y = T(h(x\alpha, \epsilon)), \quad \epsilon \sim G(.) \tag{9.3}$$

where ϵ is a real valued random variable with distribution function $G(.)$ independent of x, $h(.,.)$ is a strictly monotonic function in its second argument and $T(.)$ is a (not necessarily strictly) monotonic function. A transformation from (9.1) to the representation (9.3) may be accomplished by choosing $T(h(.,.))$ to be the generalised inverse of the conditional distribution function in (9.1) and $G(.)$ to be the uniform distribution on $(0,1)$. If (9.3) holds, (9.1) follows because the independence of ϵ and x implies that the conditional distribution of Y given $X = x$ only depends on $x\alpha$. Note that the representation (9.3) is not unique: The functions $T(.)$ and $h(.,.)$ can always be merged into a single $h^*(.,.)$ and $G(.)$ may be chosen as the uniform distribution by accommodating the function $h^*(.,.)$ accordingly. However, (9.3) allows an interpretation in terms of a latent variable $Y^* = h(x\alpha, \epsilon)$ that is observed only after some form of transformation or information loss represented by the function $Y = T(Y^*)$. Most of the "limited dependent variable" models are motivated by such a construction. Moreover, if a continuous covariate x_i with non-zero regression coefficient α_i is present, and $h(.,.)$ is also strictly monotonic in its first argument, then a conceptually simple estimator consistent for $\alpha/\|\alpha\|$ is easily constructed: take the normalised value $\hat{\alpha}$ that maximises the rank correlation between $x\alpha$ and Y (see Han (1987) for details). Obviously this works even if $T(.)$ and $h(.,.)$ are unknown. Note that the existence of consistent estimators implies identifiability as in (9.2) for that class of models (see e.g. Deistler and Seifert 1978).

To complete the description of the stochastic setup let $(Y_i, X_i)_{i=1\ldots n}$ be a sequence of i.i.d. vectors where the conditional density of Y given $X = x$ is given by (9.1), and the marginal distribution of X is non-degenerate with $\mathbb{E}(X'X) = \Sigma > 0$ and—without loss of generality—$\mathbb{E}(X) = 0$. Sometimes further assumptions on the existence of moments, the differentiability of certain functions and the interchange of differentiation and integration will be needed. These will be clear from the context.

Only maximum likelihood type estimators will be considered explicitly. Denote by

$$\sum_i \ell(y_i, x_i; \beta) = \sum_i \ln g(y_i \mid x_i\beta) \tag{9.4}$$

the log-likelihood function under an assumed model $g(. \mid .)$ in the class of general regression models but not necessarily identical to the data generating model (9.1). The first order conditions for a maximum of the log-likelihood function are given by setting the pseudo score functions to 0:

$$\sum_i U(y_i, x_i; \hat{\beta}) = \sum_i \frac{\partial}{\partial \beta} \ell(y_i, x_i; \beta)|_{\beta=\hat{\beta}} = \sum_i x_i' g^*(y_i, x_i\hat{\beta}) = 0 \quad (9.5)$$

where $g^*(.,.)$ denotes the derivative of $\ln g(. \mid .)$ with respect to its second argument. The formulation allows for nuisance parameters as long as the solution of (9.5) is unaffected by them. The sequence of estimators $\hat{\beta}_n$, $n \to \infty$ solving (9.5) will be Fisher consistent for the solution of

$$\begin{aligned} \mathbb{E}\left(U(Y,X;\hat{\beta})\right) &= \mathbb{E}\left(X'g^*(Y,X\hat{\beta})\right) \\ &= \mathbb{E}_X\left(X'\,\mathbb{E}_{Y|X}\left(g^*(Y,x\hat{\beta}) \mid X = x\right)\right) = 0 \end{aligned} \quad (9.6)$$

provided such a solution exists. This generally implies strong consistency. Under some further smoothness assumptions on $\ell(.)$, asymptotic normality of the estimators follows. See (Li and Duan 1989: 1029pp) and the literature on the behaviour of maximum likelihood type estimators in mis-specified models (see e.g. Huber 1967, White 1982, 1983, Gourieroux et al. 1984, Fahrmeir 1990). Hjort (1992) gave similar results for censored data using the counting process framework.

Since most of the arguments below use only the basic structure $\mathbb{E}(X'g^*(.,.))$ or similar structures for the expected value of the log-likelihood $\ell(.)$ under the true model, the results will also pertain to any estimators of analogous form. This includes e.g. Cox' partial likelihood in the proportional hazards case, and any M-estimator derived from maximising a criterion function by setting its derivative with respect to β to 0.

9.4. Two Special Cases: Linear Models and Least Squares under Information Loss

9.4.1. Linear Models

The "true" model is given by a location-scale model with a fixed density $f_0(.)$. That is

$$f(y \mid \alpha_0 + x\alpha) = \sigma^{-1} f_0 \left(\frac{y - \alpha_0 - x\alpha}{\sigma} \right), \; Y = \alpha_0 + x\alpha + \sigma\epsilon, \; \epsilon \sim F_0$$

Estimation is carried out by assuming another location-scale model with density $g_0(.)$:

$$g(y \mid \beta_0 + x\beta) = \tau^{-1} g_0 \left(\frac{y - \beta_0 - x\beta}{\tau} \right) \tag{9.7}$$

If maximum likelihood estimation based on the model (9.7) is used, the expected score function $U(.)$ for β becomes

$$\mathbb{E}(U(Y, X; \beta)) = -\frac{1}{\tau} \mathbb{E}_X \left(X' \mathbb{E}_{Y|X} \left[g_0^* \left(\frac{Y - \beta_0 - x\beta}{\tau} \right) \mid X = x \right] \right) = 0$$

Conditional on $X = x$

$$Y - \beta_0 - x\beta = x(\alpha - \beta) + (\alpha_0 - \beta_0) + \sigma\epsilon \tag{9.8}$$

Setting $\beta = \alpha$ makes the inner expectation independent of X so that the expectation with respect to X will be 0 (remember $\mathbb{E}(X') = 0$). Therefore $\beta = \alpha$ is a solution of the expected score function which implies consistency of the regression coefficients despite mis-specification of the disturbance distribution. This is the result of Gould and Lawless (1988): If a location-scale model is true, then using any other location-scale model for maximum likelihood estimation will lead to consistent estimators of the regression slopes, without any further restrictions on the covariates. Note that the nuisance parameters τ and α_0 do not affect the solution of the expected score equation.

9.4.2. Least Squares under Transformations and Information Loss

Assume that a general regression model holds for a latent variable Y^* with fixed distribution of the disturbance ϵ. Let $Y = T(Y^*)$ be observed, where $T(.)$ is some arbitrary function. Examples include binary regression models with $T(z) = \mathbb{1}[z \geq 0]$ (where $\mathbb{1}[.]$ is the indicator function), grouped data situations, and transformation of the dependent variable as in Box-Cox transforms. The least squares normal equations with expectation

$$\mathbb{E}(X'(Y - X\hat{\beta} - \hat{\beta}_0)) = 0$$

are used as estimating equations. The solution for the slopes is

$$\hat{\beta} = \Sigma^{-1}\mathbb{E}(X'Y)$$

Now assume that the conditional expectation of X given Y^* is linear in Y^*:

$$\mathbb{E}(X \mid Y^* = y^*) = \delta(y^* - \mathbb{E}(Y^*)) \tag{9.9}$$

This places some restrictions on the distribution of the regressors X. E.g., the condition is implied by joint normality of the regressors together with the regression structure given by (9.1).

Observe that δ is the least squares projection of X on Y^*,

$$\delta = \mathbb{E}(X'Y^*)/\text{var}(Y^*)$$

Then

$$\begin{aligned}\mathbb{E}(X'Y) &= \mathbb{E}_{Y^*}\left(Y\mathbb{E}_{X\mid Y^*}(X \mid Y^*)\right)\\ &= \mathbb{E}_{Y^*}\left(Y\delta(Y^* - \mathbb{E}(Y^*))\right) = \delta \,\text{cov}(Y, Y^*)\end{aligned}$$

By the preceding example, the regression slopes in the underlying model are given by $\alpha = \Sigma^{-1}\mathbb{E}(X, Y^*)$. Therefore

$$\begin{aligned}\beta &= \Sigma^{-1}\mathbb{E}(X'Y) = \Sigma^{-1}\delta \,\text{cov}(Y, Y^*)\\ &= \Sigma^{-1}\left(\mathbb{E}(X'Y^*)/\text{var}(Y^*)\right)\,\text{cov}(Y, Y^*)\end{aligned}$$

$$= \Sigma^{-1}\left(\Sigma\alpha/\mathrm{var}(Y^*)\right)\ \mathrm{cov}(Y,Y^*) = \alpha\left(\mathrm{cov}(Y,Y^*)/\mathrm{var}(Y^*)\right) \tag{9.10}$$

The least squares estimator in the model under information loss is proportional to the underlying vector of regression slopes and the proportionality factor is given by the least squares regression slope of Y on Y^*.

A special case of the result was given in Stewart (1983) for grouped normal regression. The present version is due to Chung and Goldberger (1984). They also point out that it is sufficient to treat the regression slopes of the least squares regression of Y^* on X as the "true" parameter α without assuming an explicit model for the conditional distribution of Y^* given X. Thus, assuming only the existence of moments, a regression result is compared with the one that would have been obtained if there had been no information loss. This operationalistic view may prove to be very helpful for judging empirical findings.

9.5. Skinner's Approach

Returning to the general situation, assume that data are generated by a model of the form (9.1) for some $f(.)$, α_0, α. Score functions from an assumed model $g(.,.)$ of the form (9.5) are used for estimation. These estimators will converge to the solution of the expected score equation (9.6). If a solution exists and is unique, Skinner (1987) proposed to use a Taylor approximation to $\beta(\alpha)$ in terms of α around $\alpha = 0$.

In fact, if $\alpha = 0$, setting $\beta = 0$ makes $g^*(.,.)$ independent of X so that the outer expectation in (9.6)—the covariance between X and $\mathbb{E}_{Y|X}(g^*(Y, x\beta) \mid X = x)$—vanishes. So a solution of (9.6) exists for $\alpha = 0$.

Taking derivatives with respect to β in the expected score equation (9.6) at $\alpha = \beta = 0$—assuming the interchange of derivative and expectation is justified—results in

$$\frac{\partial}{\partial\beta}\mathbb{E}(X'g^*(Y,X\beta))\Big|_{\alpha=\beta=0} = \mathbb{E}(X'Xg_2^*(Y,0))$$

where $g_2^*(y,\eta)$ is the derivative of $g^*(y,\eta)$ with respect to η. Since $-g_2^*(y,\eta)$ is the "observed" information for η in the assumed model, it will be positive unless the model is unidentified. Therefore its expectation with respect to Y given X will also be positive so that the derivative reduces to

$$\frac{\partial}{\partial\beta}\mathbb{E}(U(Y,X;\beta)) = \kappa\Sigma \tag{9.11}$$

for some scalar $\kappa < 0$.

Taken together, the implicit function theorem may be invoked and the function $\beta(\alpha)$ may be approximated by

$$\beta(\alpha) \simeq \beta(0) + D\beta(0)\alpha = D\beta(0)\alpha$$

around $\alpha = 0$, where D is the differential operator. It remains to compute $D\beta(0)$ explicitly. Set

$$k(x\alpha,x\beta) \coloneqq \mathbb{E}_{Y|X}(g^*(Y,x\beta) \mid X = x)$$

Then

$$\begin{aligned}
D\beta(0) &= \frac{\partial\beta(\alpha)}{\partial\alpha}\Big|_{\alpha=0} \\
&= -\left[\frac{\partial}{\partial\beta}\mathbb{E}_X(X'k(X\alpha,X\beta))\Big|_{\alpha=\beta=0}\right]^{-1} \times \\
&\qquad \left[\frac{\partial}{\partial\alpha}\mathbb{E}_X(X'k(X\alpha,X\beta))\Big|_{\alpha=\beta=0}\right] \\
&= -\left[k_2(0,0)\Sigma\right]^{-1}\left[k_1(0,0)\Sigma\right] \\
&= -\frac{k_1(0,0)}{k_2(0,0)}I = \gamma I
\end{aligned} \tag{9.12}$$

where $k_1(\mu,\eta)$, $k_2(\mu,\eta)$ are the partial derivatives of $k(\mu,\eta)$ with respect to μ resp. η and interchanges of integral and derivative are once again assumed to be valid. Since the terms for the covariances of X cancel, $D\beta(0)$ reduces to a scalar times the identity matrix. Thus the slopes of the regression under the assumed model are (to first order) proportional to the regression slopes under the data generating model.

It is instructive to compare this result to the example 9.4.2. When the linearity condition (9.9) does not hold, the conclusion of the example may not be applicable. Nevertheless, it is possible, with the help of the preceding formula to give an approximation to the least squares estimator under information loss. Specifically, let $T(z) = \mathbb{1}[z \geq 0]$. That is, least squares method is applied to binary data generated from $T(x\alpha + \alpha_0 + \epsilon)$, $\epsilon \sim F(.)$. Then:

$$\frac{\partial}{\partial\beta}\mathbb{E}_X(X'\mathbb{E}_{Y|X}(T(x\alpha + \alpha_0 + \epsilon) - x\beta - \beta_0)) = -\mathbb{E}_X(X'X) = -\Sigma$$

Moreover,

$$\begin{aligned}&\frac{\partial}{\partial\alpha}\mathbb{E}_X(X'\mathbb{E}_{Y|X}(T(x\alpha + \alpha_0 + \epsilon) - x\beta - \beta_0))\Big|_{\alpha=0}\\ &\qquad = \frac{\partial}{\partial\alpha}\mathbb{E}_X(X'(1 - F(-X\alpha - \alpha_0)))\Big|_{\alpha=0} = f(-\alpha_0)\Sigma \qquad (9.13)\end{aligned}$$

Therefore, the constant of proportionality in this case is given by

$$\gamma = f(-\alpha_0)$$

If (X,ϵ) is jointly normal, this coincides with the solution (9.10). Further explicit expressions for the constant of proportionality in special cases are given in Solomon (1984) and Galler and Pötter (1992).

Higher orders of approximation may be obtained from the implicit function theorem together with a further application of the chain rule along the lines of (9.12). To second order, the Taylor approximation is given by

$$\beta(\alpha) \simeq D\beta(0)\alpha + \frac{1}{2}(\alpha' D_l^2\beta(0)\alpha)_{l=1\ldots p}$$

with $D_l^2\beta(0)$ the Hessian of $\beta_l(\alpha)\mid_{\alpha=0}$. Now

$$\begin{aligned}D^2\beta(0) = -&\Big[\mathbb{E}(X'Xg_2^*(Y,0))\Big]^{-1}\\ &\times\Big(D_1^2\mathbb{E}_X(X'k(X\alpha,X\beta))\\ &+2D\beta(0)D_{1,2}^2\mathbb{E}_X(X'k(X\alpha,X\beta))\\ &+(D\beta(0))^2D_2^2\mathbb{E}_X(X'k(X\alpha,X\beta))\Big)\end{aligned}$$

where D_1^2 is the second order differential operator with respect to α, D_2^2 the one with respect to β and $D_{1,2}^2$ the one with respect to α and β. At $\alpha = \beta = 0$ all terms except the first (inverse) term evaluate to $\mathbb{E}(X_iX_jX_l)_{i,j,l\in\{1\ldots p\}}$ times a constant. Therefore, the second order term vanishes if all third order moments vanish. This happens e.g. if the joint distribution of X is symmetric with respect to 0, and the third moments exist. Note that no assumption of absolute continuity of the distribution of the covariates needs to be imposed. Some explicit results on the second order terms were given by Solomon (1984) in the special case of the partial likelihood under accelerated failure models.

9.6. The Approach Based on Normal Regressors

The second general approach to explain the stability of regression ratios starts with placing restrictions on the distribution of regressors. Let the expected log-likelihood $\mathbb{E}(\ell(Y;\beta_0+X\beta))$ be convex in $X\beta$. If $\mathbb{E}_{X|X\alpha}(X\beta \mid x\alpha)$ is linear in $x\alpha$ for all $\beta\in\mathbb{R}^p$, then the maximiser of the likelihood $\hat{\beta}$ converges to $\gamma\alpha$, where γ is a scalar (Li and Duan 1989, Theorem 2.1). Since α is unknown in applications, the linearity condition will normally be required for all values of α. In this case the proportionality of regression coefficients holds for all values of α, not only approximately for small α. But the restriction on the marginal distribution of the regressors is rather severe: It is e.g. implied by elliptical distributions, but discrete covariates or non-elliptical distributions are excluded.

The proof of Li and Duan's result is very simple:

$$\begin{aligned}
&\mathbb{E}\left(\ell(Y;b_0+Xb)\right)\\
&\quad=\mathbb{E}_{X\alpha,\epsilon}\left(\mathbb{E}_{X|X\alpha,\epsilon}\left(\ell\left(T(h(X\alpha,\epsilon));b_0+Xb\right)\mid X\alpha,\epsilon\right)\right)\\
&\quad\leq\mathbb{E}_{X\alpha,\epsilon}\left(\ell\left(T(h(X\alpha,\epsilon));b_0+\mathbb{E}_{X|X\alpha,\epsilon}\left(Xb\mid X\alpha,\epsilon\right)\right)\right)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\text{(by Jensen's inequality)}\\
&\quad=\mathbb{E}_{X\alpha,\epsilon}\left(\ell(T\left(h(X\alpha,\epsilon)\right);b_0+c+\gamma X\alpha)\right)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(by linearity)}
\end{aligned}$$

Therefore the expected log-likelihood at some b is always smaller than the expected log-likelihood at $\gamma\alpha$, and the result follows.

The convexity condition can be dispensed with at the price of strengthening the distributional assumptions. If X is jointly normal, then the conclusion of the theorem holds without the convexity condition. Li and Duan (1989: Theorem 2.2) prove this by a direct appeal to properties of the multivariate normal distribution. However, it is possible to prove proportionality of regression results using Stein's lemma, thus providing some insight into a possible connexion with the previous results. Recall Stein's lemma: if X is multivariate normal with covariance I (for simplicity) and $g(.)$ a real, differentiable function with $\mathbb{E}(\|Dg(X)\|)<\infty$, then $\mathbb{E}(X'g(X))=\mathbb{E}(Dg(X))$ (see e.g. Ibragimov and Has'minskii 1981: 25). Now suppose the regressors are multivariate normal. Then

$$\begin{aligned}
0=\mathbb{E}(U(X\beta,Y))&=\mathbb{E}_X(X'k(X\alpha,X\beta))\\
&=(\alpha,\beta)\mathbb{E}_X\left(\frac{k_1(X\alpha,X\beta)}{k_2(X\alpha,X\beta)}\right)
\end{aligned}$$

by Stein's lemma. Therefore,

$$\beta=-\frac{\mathbb{E}_X(k_1(X\alpha,X\beta))}{\mathbb{E}_X(k_2(X\alpha,X\beta))}\alpha \tag{9.14}$$

This proves proportionality without using the convexity of the log-likelihood, but requires normality. Note that the constant of proportionality agrees with the previous result (9.12) for $\alpha=\beta=0$, so that a Taylor argument would lead to the result of Skinner, though using much stronger assumptions. Note also that the proof starts with the expected score function, not with the expected log-likelihood. These similarities together with the second order result (9.13) prompt for generalisations either via generalisations of Stein's lemma or via Skinner's approach. But trying higher orders of the approximation seems to be less promising: although the accumulating moment conditions may lead to approximations to the normal distribution, it would require increasingly stronger differentiability assumptions for the score function, thus excluding many interesting estimators. Moreover, such assumptions are absent in the approach based on normal regressors.

Another strategy might be to apply generalisations of Stein's lemma to non-normal distributions, perhaps on the lines of Cacoullos and Papathanasiou (1992). They give identities similar to Stein's that hold for a larger class of densities, so that a standard integration by parts argument works. Unfortunately, they also show that the class of densities where this works also has to fulfil a moment condition: The conditional expectation of X_i given all other variables is linear in these variables. This is rather close to the conditions imposed by Li and Duan.

9.7. Conclusions

Under a wide variety of circumstances regression models seem to possess a remarkable property of stability: the relative magnitude of the regression coefficients are stable across different models used for estimation and under many forms of information loss. Despite of the many empirical findings a solid theoretical understanding of this property is still missing. I have concentrated here on one specific approach towards an explanation: the use of Taylor approximations via expected estimating equations. This approach has been implicit in the work of Solomon (1984), Skinner (1987) and others. I developed a more explicit version, computed the constant of proportionality and gave a second order approximation. This, and a reanalysis of the other general approach via normality assumptions showed some similarities between them. But still neither could be reduced to the other. Moreover, as the examples in section

9.4 demonstrate, neither approach is able to incorporate sufficiently these special results: in the first example the constant of proportionality equals one throughout the parameter space without any conditions on the regressors, in the second linearity of the inverse regression is needed. This is implied by the assumption of normal errors together with the linear form of the models, but not vice versa. Moreover, it was possible to compute the constant of proportionality having a nice interpretation, while the computation from the approximation formula is generally awkward.

While it seems relatively straightforward to extend both approaches to include stochastic forms of data loss (censoring, selection etc.) and to give results for multivariate response models, general results on the set of models and situations in which a given model leads to consistent estimators (up to scale) seem not to be easily derived. The work of Lazrieva and Toronjadze (1991) gives some answers in this direction, though they deal with the partial likelihood scheme which lends itself easily to such questions.

9.8. Postscriptum

While there has been no progress in developing a general theory of misspecified regression models and in particular of misspecified regression incorporating incomplete data, there is a wealth of additional results concerning particular models.

Hattori (2006) investigates an additive hazards model and shows that tests of zero effects of covariates are consistent despite misspecification of the covariate model. Angrist et al. (2006) consider quantile regression estimators and use a representation in terms of a weighted mean squared error loss function to derive an omitted variable bias formula. Müller et al. (2008) investigate the effects of estimators of semi-Markov models when transition distributions are misspecified. Müller (2007) considers misspecified nonlinear regression models. O'Brien et al. (2006) consider the effect of misspecification on the power of tests of association of covariates with a dependent variable in the class of generalised linear models. Pantazis and Touloumi (2007) present a simulation study on the robustness of parametric models for bivariate censored data when the censoring is informative and the joint distribution is misspecified. Zhang et al. (2006) discuss the effects of misspecified linear transformation models in the presence of censoring. Pascual (2005) derives asymptotic approximations to the bias of parameter estimates in misspecified log-normal and Weibull distributions when data can either be type I or type II censored. Fushiki (2005) treats the statistical prediction problem when models are misspecified. Similarly, Cai et al. (2008) consider predictions based on misspecified regression models. Meister (2004) shows the effect of misspecifying the error density in a deconvolution problem on the mean integrated squared error. Gustafson (2001) suggests sensitivity measures for the degree of misspecification in parametric models. Finally, DiRienzo and Lagakos (2001) analyse the effect of model specification on the power of tests of no effect in censored Cox models.

Most investigations of misspecification in recent years concentrated on the analysis of misspecified dependence structures in longitudinal data models. I mention only Keiding et al. (1997), Litière et al. (2007), Jowaheer (2006), Agresti et al. (2004), Wang and Carey (2003), Rizopoulos et al. (2008), Wan et al. (2007), Wang and Lin (2005), Verbeke and Fieuws (2007), Jacqmin-Gadda et al. (2007), Cheng and Shao (2006).

A more general approach has been advocated by Royall and Tsou (2003) who construct robust adjusted likelihood ratios in general parametric models (see also Kateri and Balakrishnan (2008) for an application to contingency tables). Patilea (2001) investigates the connection between convexly parametrised models and the behaviour of maximum likelihood estimators when the model is misspecified. Classes of convexly parametrised models are also used by van der Laan and Robins (2003) in their discussion of the double robustness of certain estimators in incomplete data models. Finally, Brown et al. (2006) survey recent results and generalisations of Stein's Lemma.

A

An Implementation of a Class of Distribution Function Estimators

This section describes an implementation of a basic building block in many incomplete data problems, namely the construction of estimators of a cumulative distribution function from right censored data. The building block has been used for the multivariate Buckley-James estimator suggested in Chapter 8 and is closely related to the estimation method used in Chapter 6. I will start by describing the general setup of the software and will then discuss the particular algorithms used in a variety of special cases.

The procedure is implemented in `TDA` (available from `http://www.stat.rub.de/tda.html`). A command `gdf` is provided that can be used to calculate marginal and joint distribution and survivor functions based on possibly incomplete (censored) data. The syntax of the command is shown in Box 1. Most parameters are optional. Required is the name of a variable (specified with the `yl` parameter) and the name of an output file to be given on the right-hand side.

Input data must be defined for $i = 1, \ldots, n$ units. Each unit can contribute observations for (a subset of) m dimensions. If available, y_{ij} is the observation for unit i in dimension j. In general, each observation consists of two parts

$$(y_{ij}, \delta_{ij})$$

where y_{ij} is the observed value, and δ_{ij} indicates whether the observation is uncensored ($\delta_{ij} = 1$), or is right censored ($\delta_{ij} = 0$). The corresponding data matrix variables can be specified with the `yl` and `cen` parameters, respectively. The following combinations are possible.

1. Only observed values are specified with the `yl` parameter. Then all observations are assumed to be exact with corresponding values.
2. Observed values are specified with the `yl` parameter and a censoring indicator is specified with the `cen` parameter. An observations is then interpreted as exact at y_{ij} if $\delta_{ij} = 1$ and is interpreted as right censored if $\delta_{ij} = 0$.

By default, the command assumes a single dimension ($m = 1$) and $n =$ `NOC` units, where `NOC` is the number of cases in the current data matrix. The `grp` parameter can be used to specify multivariate data. The syntax is

```
grp = ID, L1,
```

where `ID` and `L1` are names of data matrix variables. Each block of data matrix rows where `ID` has identical values is interpreted as data for one unit. Any values are possible and since the data matrix is always sorted with respect to `ID` and `L1`, it is not required that blocks are contiguous. The `L1` variable must contain positive integers. The number of different integers found in this variable is interpreted as the number dimensions. Again, it is not required that these numbers are contiguous. Each data matrix row provides one observation for the unit given by the `ID` variable and dimension given by the corresponding `L1` variable.

To illustrate, consider the data in Box 2. In this example there are three units and six observations in all. The number of dimensions is $m = 3$. Dimensions are mapped to the values of the `L1` variable in ascending order. Thus, dimension 1, 2, and 3 correspond, respectively, to `L1 = 1`, `L1 = 3`, and `L1 = 4`. The first unit (`ID = 1`) has observations for

Box 1: Syntax for `gdf` command.

```
gdf (
   opt=...,       method, def. 1
                  1 = marginal calculation
                  2 = joint calculation, method 1
                  3 = joint calculation, method 2
   prn=...,       output option, def. 0
                  0 = distribution functions
                  1 = survivor functions
                  2 = expected values
   yl=...,        variable name for observed values
   cen=...,       variable name for censoring information
   grp=ID,L1,     specification of dimensions

   sc=...,        offset for domain (method 1 and 2), def. 0
   n=...,         number of boxes in grid (method 1), def. 100
   mxit=...,      maximal number of iterations (method 1), def. 20
   tolf=...,      tolerance for convergence (method 1), def. 0.001
   d=...,         delta specification (method 2), def. 0.1
   fmt=...,       print format for output file, def. 10.4
   prot=...,      protocol file with diagnostic information
) = fname;
```

Box 2: Example data to illustrate `grp` parameter.

```
 ID  L1  YL   D
---------------
  1   1   3   1
  1   3   5   1
  1   4   4   0
  2   3   0   1
  5   1   7   0
  2   4   1   0
```

all three dimensions. The observation is exact in the first and second dimension, and right censored in the third dimension. Unit 2 contributes an exact observation to the second dimension and a right censored observation to the third dimension. The third unit (`ID = 5`) contributes a right censored observation to the first dimension.

A.1. Marginal Distribution Functions

If `opt = 1` (default), the command calculates marginal distributions, separately for each dimension present in the input data. Assuming that there are n_j observations for the jth dimension, the data are:

$$(y_{1j}, \delta_{1j}), \dots, (y_{n_j j}, \delta_{n_j j})$$

Depending on `prn`, these data are used to calculate a distribution function (`prn=0`), a survivor function (`prn=1`), or expected values (`prn=2`).

Marginal EDF: Exact Observations

If the data for one dimension contain only exact observations, the command calculates a standard empirical distribution function,

$$F_j(y_{ij}) = \sum_{i=1}^{n_j} \mathbb{1}(Y_j \leq y_{ij})$$

if `prn = 0`, or survivor function,

$$S_j(y_{ij}) = 1 - F_j(y_{ij})$$

if `prn = 1`. Here Y_j denotes the variable in the jth dimension and refers to the possible values y_{ij}, $i = 1, \dots, n_j$. $\mathbb{1}()$ denotes the indicator function. In this case, if `prn = 2`, the expected values equal the observed values.

Box 3 provides an illustration. Input data are given by the variable `YL`. The command

```
gdf (yl=YL) = df0;
```

Box 3: Illustration of calculations with exact observations.

```
             df0 (prn = 0)                  ds0 (prn = 1)
   YL  D        Y         F          D       Y          S
   --  -----------------------    -----------------------
    1  1    -1.0000    0.1667    1    -1.0000     0.8333
    7  1     1.0000    0.3333    1     1.0000     0.6667
   -1  1     5.0000    0.5000    1     5.0000     0.5000
    5  1     7.0000    0.8333    1     7.0000     0.1667
    7  1     9.0000    1.0000    1     9.0000     0.0000
    9
```

creates the output file df0, the command

```
gdf (yl=YL,prn=1) = ds0;
```

creates the output file ds0.[1] The first column in the output files shows the dimension, then follow the value of the variable (sorted in ascending order) and the corresponding value of the distribution or survivor function. If the input data refer to more than one dimension, the output file will contain the same information separately for each dimension.

Marginal EDF: Right Censored Observations

A second situation occurs if the input data, for one dimension, contain both, exact and right censored observations. The command then uses the standard Kaplan-Meier procedure to calculate a marginal distribution, or survivor, function. Considering the jth dimension, the observations for Y_j are sorted in ascending order. If there are exact and right censored observations for the same value of the variable, exact observations come first, followed by right censored observations. The highest value of Y_j is always treated as uncensored. In consequence, the expectation computed from the estimated distribution function will always be finite. Then, in the order from lowest to highest values, the mass of each right censored observation is distributed to the right (over all observations with larger values, see Efron (1988)).

[1] The data file for this and the following examples is gdf1.dat. The command file is gdf1.cf. Both are contained in the TDA example archive.

Box 4: Marginal distributions with right censored observations.

```
                     df1 (prn = 0)                 ds1 (prn = 1)

 YL DELTA   D      Y          F          D      Y          S
 --------   -----------------------     -----------------------
  3    1    1      1.0000     0.1667     1      1.0000     0.8333
  3    0    1      3.0000     0.3750     1      3.0000     0.6250
  2    0    1      4.0000     0.6875     1      4.0000     0.3125
  1    1    1      5.0000     1.0000     1      5.0000     0.0000
  4    1
  5    1
```

An example is given in Box 4. Input data are given by the variables `YL` and `DELTA`. The command

```
gdf (yl=YL,cen=DELTA) = df1;
```

computes the distribution function. Adding the parameter `prn = 1` computes the corresponding survivor function.

A.1.1. Marginal EDF: Expected Values

If `prn = 2`, the `gdf` command calculates expected values based on the marginal distributions. If an observation is exact its expected value equals its observed value. If the observation is right censored, the command calculates

$$y_{ij}^{*} = \mathbb{E}_{F_j}(Y_j \mid Y_j > y_{ij}) = \int_{y_{ij}}^{\infty} y \, dF_j \Big/ \int_{y_{ij}}^{\infty} dF_j$$

where F_j is the Kaplan-Meier estimate of the marginal distribution function in the jth dimension.

To illustrate, consider the data in Box 5 (same as in Box 4). The command is now

```
gdf (yl=YL,cen=DELTA,prn=2) = de1;
```

Box 5: Expected values based on marginal distribution.

```
                 de1 (prn = 2)

 YL DELTA    D       Y      CEN    E
 --------   --------------------------
  3   1      1      3.0000  1    3.0000
  3   0      1      3.0000  0    4.5000
  2   0      1      2.0000  0    4.1250
  1   1      1      1.0000  1    1.0000
  4   1      1      4.0000  1    4.0000
  5   1      1      5.0000  1    5.0000
```

The resulting output file, also shown in Box 5, contains four columns. The first column refers to the current dimension, the second column contains the observed values, and the third column shows the censoring status of the observation. The last column contains the expected value. This will equal the observed value if the observation is uncensored, otherwise the expected value as calculated from the Kaplan-Meier survivor function. Note that with this option, the input data are not sorted. Note also that the largest observation is always assumed to be uncensored.

A.2. Joint Distributions

We now discuss the calculation of joint distribution functions. This is done separately for each marginal pattern which is found in the input data. The word "“marginal pattern"” refers to a combination of dimensions,

$$(d_1, ..., d_{m_k}) \qquad \text{where} \qquad 1 \leq d_1 < d_2 < \cdots < d_{m_k} \leq m$$

m being the maximal number of dimensions as given by the input data. For ease of notation, the following discussion refers to a full marginal pattern, i.e. that $m_k = m$. Now, let n denote the number of units for the marginal pattern. We then have m observations for each unit, as follows:

$$(y_{i1}, \delta_{i1}), \ldots, (y_{im}, \delta_{im}) \qquad i = 1, \ldots, n$$

Based on these data, the command calculates marginal distributions (if `opt = 1`) or joint distributions (if `opt = 2` or `opt = 3`). Calculations depend on whether the data contain right censored observations. The contents of the resulting output file depends on the `prn` parameter. The command calculates a distribution function if `prn = 0`, a survivor function if `prn = 1`, or expected values if `prn = 2`.

A.2.1. Joint Distribution: Exact Observations

If all observations are exact, the command calculates a standard m-dimensional distribution function, defined as

$$F(y_1, \ldots, y_m) = \sum_{i=1}^{n} \prod_{j=1}^{m} \mathbb{1}(Y_j \leq y_i)$$

or the corresponding survivor function

$$S(y_1, \ldots, y_m) = \sum_{i=1}^{n} \prod_{j=1}^{m} \mathbb{1}(Y_j > y_i)$$

The functions are calculated and tabulated in the output file, for all data points in the input data.

To illustrate, consider the data shown in Box 6. There are four units, all having observations for two dimensions. The joint distribution function, shown in the upper half of the right part of the box, was calculated with the command

```
gdf (grp=ID,L1,yl=YL,opt=2,prn=0) = df3;
```

The corresponding survivor function was calculated with the command

```
gdf (grp=ID,L1,yl=YL,opt=2,prn=1) = ds3;
```

In both cases, the first column in the output file refers to the current marginal pattern, followed by the dimensions (values of `L1` variable) that define this pattern. The next m columns contain the observations, Box 6: Joint distribution with exact observations

```
 ID  L1    YL      MP  D1 D2      F          Y1          Y2
------------      ---------------------------------------------
  1   1   1.0       1   1  2    0.2500    -1.0000      1.0000
  1   2   2.0       1   1  2    0.5000     1.0000      2.0000
  2   1  -1.0       1   1  2    0.5000     2.0000      1.0000
  2   2   1.0       1   1  2    0.7500     3.0000      1.5000
  3   1   3.0
  3   2   1.5      MP  D1 D2      S          Y1          Y2
  4   1   2.0     ---------------------------------------------
  4   2   1.0       1   1  2    0.0000     3.0000      1.5000
                    1   1  2    0.2500     2.0000      1.0000
                    1   1  2    0.0000     1.0000      2.0000
                    1   1  2    0.5000    -1.0000      1.0000
```

sorted in ascending or descending order. The final column contains the corresponding value of the distribution or survivor function.

If there are only exact observations, and `prn = 2`, the output file will simply contain the observations, a censoring indicator that always has value 1, and expected values that equal the observed values. In this case, the input data are not sorted.

A.2.2. Censored Observations, Method 1

We now consider a situation where the multivariate data contain right censored observations. Unfortunately, there is no simple generalization of the 1-dimensional Kaplan-Meier procedure. Discussion in the literature has proposed several different approaches. The `gdf` command offers two methods. The first one (selected with `opt = 2`) can be used with observations which might be censored simultaneously in several dimensions. The second method (selected with `opt = 3`) can only be used with observations which are censored at most in one dimension. This section describes the first method (`opt = 2`).

This method uses an iterative EM-like procedure that tries to find a self-consistent estimate of the distribution function. In explaining the procedure we refer, again, to a full marginal pattern. Data are given by

$$(y_{i1}, \delta_{i1}), \ldots, (y_{im}, \delta_{im}) \qquad i = 1, \ldots, n$$

In a first step, we calculate a domain as an m-dimensional interval

$$D = D_1 \times \cdots \times D_m$$

where

$$D_j = \;]\min_i \{y_{ij}\} - \sigma, \max_i \{y_{ij}\} + \sigma]$$

By default, the offset is $\sigma = 0$. This implies that observations which have an exact component on a left side of the domain, or a right censored component on a right side of the domain, will not be used.[2] In order to include all observations one can specify a positive offset with the sc parameter, see Box 1.

The iterative procedure is based on a partition of the domain into a grid of boxes. The jth dimension is partitioned into q_j intervals

$$i_j(k) = \;]\, l_j(k), u_j(k)\,] \qquad k = 1, \ldots, q_j$$

These intervals, and the corresponding boxes, are treated as open on the left side and closed on the right side. Since the number of boxes rapidly increases in higher dimensions, we require the user to specify a total number of boxes for the whole grid with the n parameter, default is $n = 100$. The command then tries to find an integer q such that $q^m \approx n$ and sets

$$q_1 = \ldots = q_m = q$$

The minimum is $q = 1$, that is, the grid consists of only a single box. Let now $B_j = \{1, \ldots, q_j\}$. Then, each

$$(k_1, \ldots, k_m) \in B_1 \times \cdots \times B_m$$

[2] Corresponding values of the distribution function, or survivor function, will then be -1, and expected values will then equal the observed (possibly censored) values.

refers to one box in the grid, namely

$$B(k_1,\ldots,k_m) = i_1(k_1) \times \cdots \times i_m(k_m)$$

We now define for each observation (y_{ij}, δ_{ij}) a subset

$$b_j(y_{ij}, \delta_{ij}) \subseteq B_j$$

containing pointers to those boxes (in the jth dimension) where the observation possibly has values. Explicitly, if the observation is exact, the definition is

$$b_j(y_{ij}, \delta_{ij}) = \{\, k \in B_j \mid y_{ij} \in i_j(k) \,\}$$

and if the observation is right censored, the definition is

$$b_j(y_{ij}, \delta_{ij}) = \{\, k \in B_j \mid \exists\, \delta > 0 : \ y_{ij} + \delta \in i_j(k) \,\}$$

Then, for each unit i,

$$\bar{b}_i = b_1(y_{i1}, \delta_{i1}) \times \cdots \times b_m(y_{im}, \delta_{im})$$

provides pointers to those boxes in the domain where unit i has, possibly, an m-dimensional value. Of course, $\bar{b}_i$ will be empty, if unit i has a component not covered by the domain.

Using these notations, Box 7 shows the iterative algorithm.[3] The maximal number of iterations can be specified with the `mxit` parameter, default is 20. Convergence is assumed if

$$\max_{k_1,\ldots,k_m} \{\, |\, f(k_1,\ldots,k_m) - f'(k_1,\ldots,k_m) \,|\,\} \ \leq \ \text{TOLF}$$

where $f'()$ refers to the density from the previous iteration. By default, TOLF $= 0.001$; other values can be specified with the `tolf` parameter.

Information about the number of iterations and the final value of the convergence criterion is given in the standard output. In any case,

[3] The notation '+=' means that the expression on the right-hand side is added to the expression on the left-hand side.

Box 7: Iterative algorithm for joint distribution (method 1)

```
(1) $\forall\,(k_1,\dots,k_m) \in B_1 \times \cdots \times B_m \;:\; f^*(k_1,\dots,k_m) = 0$
(2) $\forall\,i\;\forall\,(k_1,\dots,k_m) \in \bar{b}_i \;:\; f^*(k_1,\dots,k_m) \mathrel{+}= \dfrac{1}{n\,|\,\bar{b}_i\,|}$
(3) $\forall\,(k_1,\dots,k_m) \in B_1 \times \cdots \times B_m \;:\; f(k_1,\dots,k_m) = 0$
(4) $\forall\,i\;\forall\,(k_1,\dots,k_m) \in \bar{b}_i \;:$
    $f(k_1,\dots,k_m) \mathrel{+}= \dfrac{1}{n}\dfrac{f^*(k_1,\dots,k_m)}{\sum_{(l_1,\dots,l_m)\in\bar{b}_i} f^*(l_1,\dots,l_m)}$
(5) end if convergence has been achieved, or the maximal
    number of iterations has been reached.
(6) $\forall\,(k_1,\dots,k_m) \in B_1 \times \cdots \times B_m \;:$
    $f^*(k_1,\dots,k_m) = f(k_1,\dots,k_m)$
(7) continue with (3)
```

depending on prn, the gdf command finally calculates a distribution function, a survivor function, or expected values. In order to explain the calculation, let

$$k_{ij} = \min\{\, k_j \mid k_j \in b_j(y_{ij}, \delta_{ij}) \,\}$$

meaning that k_{ij} refers to the box where, in dimension j, observation y_{ij} begins. The distribution function is then calculated using the formula

$$F(y_{i1},\dots,y_{im}) \;=\; \sum_{k_1=1,\dots,k_{i1}} \cdots \sum_{k_m=1,\dots,k_{im}} f(k_1,\dots,k_m)$$

Correspondingly, calculation of the survivor function uses the formula

$$S(y_{i1},\dots,y_{im}) \;=\; \sum_{k_1=k_{i1},\dots,q_1} \cdots \sum_{k_m=k_{im},\dots,q_m} f(k_1,\dots,k_m)$$

Box 8: Example data set (Pruitt [1993])

```
ID  L1   Y  CEN
---------------
 1   1   1   0
 1   2   6   1
 2   1   2   0
 2   2   4   1
 3   1   3   0
 3   2   5   0
 4   1   4   1
 4   2   3   1
 5   1   5   1
 5   2   2   0
 6   1   6   1
 6   2   7   1
 7   1   7   0
 7   2   1   0
 8   1   8   1
 8   2   8   1
```

If `prn = 2`, the command calculates expected values. To explain this option, let $(y_{i1}, \ldots, y_{im})$ be one of the m-dimensional observations. The expected value, in the jth dimension, will equal y_{ij} if this component is not censored. Otherwise, it is calculated by

$$\frac{\sum_{(k_1,\ldots,k_m)\in\bar{b}_i} z_j(k_1,\ldots,k_m)\,f(k_1,\ldots,k_m)}{\sum_{(k_1,\ldots,k_m)\in\bar{b}_i} f(k_1,\ldots,k_m)}$$

$z_j(k_1,\ldots,k_m)$ is the jth component of the mean value of all exact observations falling in box $(k_1,\ldots,k_m)$ or, if the box does not contain exact observations, equals the mean of $i_j(k_j)$.

Example 1. For a first illustration we use some example data from Pruitt [1993], shown in Box 8.[4] In order to estimate a distribution function, we use the command

```
gdf(opt=2,prn=0,yl=Y,cen=CEN,grp=ID,L1,n=64,sc=0.5) = df5;
```

[4] The data file is `gdf2.dat`, contained in the TDA example archive.

Box 9: Estimated distribution function

```
   MP   D1  D2      Y1          Y2           F
  ------------------------------------------
    1    1   2      1.0000      6.0000      0.0000
    1    1   2      2.0000      4.0000      0.0000
    1    1   2      3.0000      5.0000      0.0001
    1    1   2      4.0000      3.0000      0.1250
    1    1   2      5.0000      2.0000      0.0000
    1    1   2      6.0000      7.0000      0.6290
    1    1   2      7.0000      1.0000      0.0000
    1    1   2      8.0000      8.0000      1.0000
```

Box 10: Estimated expected values

```
   MP   D1  D2     Y1 (obs)    Y2 (obs)    CEN     Y1 (est)    Y2 (est)
  ----------------------------------------------------------------
    1    1   2      1.0000      6.0000    0 1      5.6197      6.0000
    1    1   2      2.0000      4.0000    0 1      5.5839      4.0000
    1    1   2      3.0000      5.0000    0 0      6.6101      6.9849
    1    1   2      4.0000      3.0000    1 1      4.0000      3.0000
    1    1   2      5.0000      2.0000    1 0      5.0000      5.1561
    1    1   2      6.0000      7.0000    1 1      6.0000      7.0000
    1    1   2      7.0000      1.0000    0 0      7.8492      7.1540
    1    1   2      8.0000      8.0000    1 1      8.0000      8.0000
```

In this example, we have used a total number of 64 boxes and added a small offset to the domain in order to cover all observations.[5]

Box 9 shows the resulting output file. Estimated expected values, calculated with the `prn=2` option are shown in Box 10.

Example 2. For a second illustration, we create 100 data points

$$(y_{i1}, y_{i2}) \qquad i = 1, \ldots, 100$$

where

$$y_{i1} = i, \; y_{i2} = i + r_i$$

[5] The command file is `gdf2.cf`.

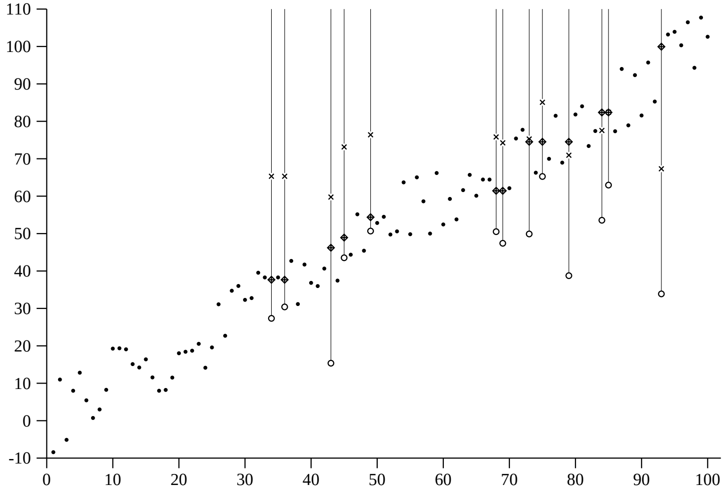


Figure A.1.: Illustration of observed data points and estimated expected values. Estimation with marginal Kaplan-Meier procedure (x) and with joint estimation (method 1, n = 100 boxes).

and r_i are random numbers which are equally distributed in $[-10, 10]$. We then have randomly censored 13 of these data points in the second dimension. The resulting data points are shown in Figure 1.

We then estimated expected values, first using a marginal Kaplan-Meier procedure.[6] The resulting estimated values are indicated in Figure 1 by cross (x) symbols. We then used the iterative procedure described above to estimate expected values. The resulting estimated values are indicated in Figure 1 by ♦ symbols. They obviously provide somewhat better estimates.

[6] The command file for data generation and estimation is `gdf3.cf` in the TDA example archive.

Box 11: Algorithm for joint distribution (method 2)

```
(1) Sort observations $(y_{i1}, \dots, y_{im})$ in ascending order,
    first wrt first component, then wrt second component,
    and so on; in case of ties, exact observations precede
    censored observations.
(2) for $i = 1, \dots, n \ : \ f(i) = 1/n$
(3) $i = 1$

(4) If $(y_{i1}, \dots, y_{im})$ is exact in all components,
    continue with step (8).

(5) Let $(y_{i1}, \dots, y_{im})$ be censored in dimension $j_0$.
    Calculate an index set $B(i, j_0)$ containing indices of all
    data points $(y_{k1}, \dots, y_{km})$ for which:
    a) $y_{kj} > y_{ij}$
    b) $\forall j \neq j_0 : \ y_{kj}$ is exact
    c) $\forall j \neq j_0 : \ |\, y_{kj} - y_{ij} \,| \ \leq \Delta_{j_0}/2$
(6) $\forall k \in B(i, j_0) : \ f(k) \mathrel{+}= \ \dfrac{f(i)}{|\, B(i, j_0) \,|}$
(7) $f(i) = 0$
(8) $i \mathrel{+}= \ 1$
(9) if $i \leq n$ continue with (4).
```

A.2.3. Censored Observations, Method 2

We now describe an alternative approach (selected with `opt = 3`) that can be used when the observations are censored in at most a single dimension. The basic idea is quite simple: we use a local Kaplan-Meier procedure based on all observations that are available in one of the censored dimensions.

Box 11 explains the algorithm. It is controlled by parameters Δ_j that define the size of the subsets of the domain used for the local Kaplan-Meier procedure. These parameters are calculated by using the parameter d that can be specified by the user, see Box 1. Then

$$\Delta_j = d\, W_j$$

where W_j denotes the width of the domain in dimension j. Default is $d = 0.1$.

The algorithm results in densities, $f(i)$, for all data points i that do not contain censored observations. Depending on `prn`, they are finally used to calculate a distribution function

$$F(y_{i1}, \ldots, y_{im}) = \sum_{(y_{k1},\ldots,y_{km}) \leq (y_{i1},\ldots,y_{im})} f(k)$$

if `prn = 0`, or a survivor function

$$S(y_{i1}, \ldots, y_{im}) = \sum_{(y_{k1},\ldots,y_{km}) > (y_{i1},\ldots,y_{im})} f(k)$$

if `prn = 1`. These values are tabulated in the output file for all data points. The data points are not sorted.

If `prn = 2`, the command calculates expected values. For each data point $(y_{i1}, \ldots, y_{im})$, if y_{ij} is exact, this will equal the corresponding expected component. If y_{ij} is censored, the expected value is calculated by

$$\hat{y}_{ij} = \frac{\sum_{k \in B(i,j)} y_{kj} f(k)}{\sum_{k \in B(i,j)} f(k)}$$

For the definition of $B(i, j)$ see Box 11. If this index set is empty, $\hat{y}_{ij}$ will equal the observed value, y_{ij}.

Example 3. For an illustration, we use the 2-dimensional data from example 2. There are 100 data points, 13 are censored in the second dimension. Since the data points are censored in only a single dimension, we can use both methods, 1 and 2. Box 12 shows the censored data points

Box 12: Estimated expected values (method 1 and method 2)

```
                                          method 1  method 2
Y1 (obs)  Y2 (obs)  CEN    Y1 (est)  Y2 (est)  Y2 (est)
--------------------------------------------------------
34.0000   27.3464   1 0    34.0000    37.6603   37.3177
36.0000   30.3996   1 0    36.0000    37.6603   38.0035
43.0000   15.3593   1 0    43.0000    46.2104   46.2348
45.0000   43.5366   1 0    45.0000    48.9264   48.4735
49.0000   50.6783   1 0    49.0000    54.3749   51.9167
68.0000   50.5112   1 0    68.0000    61.4348   71.7563
69.0000   47.4082   1 0    69.0000    61.4348   71.7563
73.0000   49.9009   1 0    73.0000    74.5214   73.0277
75.0000   65.2591   1 0    75.0000    74.5214   73.7913
79.0000   38.7495   1 0    79.0000    74.5214   79.1623
84.0000   53.5625   1 0    84.0000    82.3746   83.4225
85.0000   62.9593   1 0    85.0000    82.3746   85.3742
93.0000   33.8836   1 0    93.0000    99.9475  103.4777
```

and estimated expected values for their censored component.[7] It is seen that, in this example, both methods give quite similar estimates.

Example 4. If `d = 1`, the algorithm uses all observations (in the censored dimension) and becomes identical with a standard marginal Kaplan-Meier procedure. For an illustration see command file `gdf5.cf` in the TDA example archive.

[7] The command file is `gdf4.cf`.

Bibliography

Aakvik, Arild 2001: Bounding a matching estimator: The case of a Norwegian training program. Oxford Bulletin of Economics and Statistics 63: 115–143.

Aalen, Odd O. 1987: Dynamic modelling and causality. Scandinavian Actuarial Journal: 177–190.

Aalen, Odd O., John Fosen, Harald Weedon-Fekjaer, Ørnulf Borgan, Einar Husebye 2004: Dynamic analysis of multivariate failure time data. Biometrics 60: 764–773.

Aalen, Odd O., Arnoldo Frigessi 2007: What can statistics contribute to a causal understanding? Scandinavian Journal of Statistics 34: 155–168.

Abadie, Alberto, Guido W. Imbens 2006: Large sample properties of matching estimators for average treatment effects. Econometrica 74: 235–267.

Abayomi, Kobi, Andrew Gelman, Marc Levy 2008: Diagnostics for multivariate imputations. Applied Statistics 57: 237–291.

Adams, Peter, Michael D. Hurt, Daniel McFadden, Angela Merrill, Tiago Ribeiro 2003: Healthy, wealthy, and wise? Tests for direct causal paths between health and socioeconomic status (with discussion). Journal of Econometrics 112: 3–133.

Addison, J.T., P. Portugal 1987: On the distributional shape of unemployment duration. Review of Economics and Statistics 69: 520–526.

Aerts, Marc, Gerda Claeskens, Niel Hens, Gert Molenberghs 2002: Local multiple imputation. Biometrika 89: 375–388.

Afifi, A.A., R.M. Elashoff 1966: Missing observations in multivariate statistics I: Review of the literature. Journal of the American Statistical Association 61: 595–604.

Afifi, A.A., R.M. Elashoff 1967: Missing observations in multivariate statistics II. Point estimation in simple linear regression. Journal of the American Statistical Association 62: 10–29.

Afifi, A.A., R.M. Elashoff 1969: Missing observations in multivariate statistics III: Large sample analysis of simple linear regression. Journal of the American Statistical Association 64: 337–358.

Afifi, A.A., R.M. Elashoff 1969: Missing observations in multivariate statistics IV: A note on simple linear regression. Journal of the American Statistical Association 64: 359–365.

Agresti, Alan, Brian Caffo, Pamela Ohman-Strickland 2004: Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. Computational Statistics & Data Analysis 47: 639–653.

Ahn, Jeongyoun, J.S. Marron, Keith M. Muller, Yueh-Yun Chi 2007: The high-dimension, low-sample-size geometric representation holds under mild conditions. Biometrika 94: 760–766.

Aigner, Martin 1997^2: Combinatorial Theory. Berlin: Springer.

Aitkin, Murray 1981: A note on the regression analysis of censored data. Technometrics 23: 161–163.

Akritas, Michael G., Jouni Kuha, D. Wayne Osgood 2002: A nonparametric approach to matched pairs with missing data. Sociological Methods & Research 30: 425–454.

Akritas, Michael G., Ingrid van Keilegom 2003: Estimation of bivariate and marginal distributions with censored data. Journal of the Royal Statistical Society B 65: 457–471.

Alavi, A., A. Thavenaswaran 2002: Nonparametric estimators for censored correlated data. Communications in Statistics — Theory and Methods 31: 977–985.

Albert, Paul S., Dean A. Follmann 2003: A random effects transition model for longitudinal binary data with informative missingness. Statistica Neerlandica 57: 100–111.

Albert, Paul S., Dean A. Follmann 2007: Random effects and latent processes approaches for analyzing binary longitudinal data with missingness: A comparison of approaches using opiate clinical trial data. Statistical Methods in Medical Research 16: 417–439.

Aldrich, J. 1989: Autonomy. Oxford Economic Papers 41: 15–34.

ALLBUS 1998: Codebuch des kumulierten ALLBUS 1980–96. ZA–Nummer 1795: Release 98:01: Köln: Zentralarchiv für Empirische Sozialforschung.

Allison, Paul D. 2000: Multiple imputation for missing data. A cautionary tale. Sociological Methods & Research 28: 301–309.

Almeida, Carlos, Michel Mouchart 2007: Bayesian encompassing specification test under not completely known partial observability. Bayesian Analysis 2: 303–318.

Alvarez, M., J. Hyman 1998: Agents and their actions. Philosophy 73: 219–245.

Ambler, Gareth, Rumana Z. Omar, Patrick Royston 2007: A comparison of imputation techniques for handling missing prdictor values in a risk model with a binary outcome. Statistical Methods in Medical Research 16: 277–298.

Andersen, Elisabeth Wreford 2005: Two-stage estimation in copula models used in family studies. Lifetime Data Analysis 11: 333–350.

Andersen, Per Kragh, Ørnulf Borgan, Richard D. Gill, Niels Keiding 1993: Statistical Models Based on Counting Processes. Berlin: Springer.

Andersen, Per Kragh, Knut Liestøl 2003: Attenuation caused by infrequently updated covariates in survival analysis. Biostatistics 4, 633–649.

Andrews, Chris, Mark van der Laan, James Robins 2005: Locally efficient estimation of regression parameters using current status data. Journal of Multivariate Analysis 96: 332–351.

Andrews, Donald W.K., Marcia M.A. Schafgans 1998: Semiparametric estimation of the intercept of a sample selection model. Review of Economic Studies 65: 497–517.

Angrist, Joshua D. 1997: Conditional independence in sample selection models. Economics Letters 54: 103–112.

Angrist, Joshua D., Guido W. Imbens 1999: Comments on James J. Heckman, "Instrumental variables. A study of implicit behavioral assumptions used in making program evaluations." (with response). The Journal of Human Resources 34: 823–837.

Angrist, Joshua D., Guido W. Imbens, Donald B. Rubin 1996: Identification of causal effects using instrumental variables (with discussion). Journal of the American Statistical Association 91: 444–472.

Angrist, Joshua D., Victor Chernozhukov, Iván Fernández-Val 2006: Quantile regression under misspecification, with an application to the U.S. wage structure. Econometrica 74: 539–563.

Anscombe, Gertrude Elizabeth Margaret 1971: Causality and Determination. An Inaugural Lecture. Cambridge: Cambridge University Press.

Antony, Ansa Alphonsa, P.G. Sankaran 2008: Nonparametric estimation of bivariate survivor function under masked causes of failure. Journal of Nonparametric Statistics 20: 77–89.

Arellano-Valle, Reinaldo B., Márcia D. Branco, Marc G. Genton 2006: A unified view on skewed distribuitions arising from selections. Canadian Journal of Statistics 34: 581–601.

Arjas, Elja, M. Eerola 1993: On predictive causality in longitudinal studies. Journal of Statistical Planning and Inference 34: 361–386.

Arjas, Elja, P. Haara 1984: A marked point process approach to censored failure data with complicated covariates. Scandinavian Journal Statistics 11: 193–209.

Arnap, Raghuqnath, Darjinder Singh 2006: A new method for estimating variance from data imputed with ratio method of imputation. Statistics & Probability Letters 76: 513–519.

Asgharian, Masoud, Cyr Emile M'Lan, David B. Wolfson 2002: Length-biased sampling with right censoring: An unconditional aproach. Journal of the American Statistical Association 97: 201–209.

Asgharian, Masoud, David B. Wolfson 2005: Asymptotic behavior of the unconditional NPMLE of the length-biased survivor function from right censored prevalent cohort data. Annals of Statistics 33: 2109–2131.

Aubin, Jean-Pierre, Hélène Frankowska 1990: Set-Valued Analysis. Boston: Birkhäuser.

Austin, Peter C. 2008: A critical appraisal of propensity-score matching in the medical literature between 1996" and 2003 (with discussion). Statistics in Medicine 27: 2037–2069.

Austin, Peter C., Michael D. Escobar 2005: Bayesian modeling of missing data in clinical research. Computational Statistics & Data Analysis 49: 821–836.

Austin, Peter C., Paul Grootendorst, Sharon-Lise T. Normand, Geoffrey M. Anderson 2007a: Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: A Monte Carlo study. Statistics in Medicine 26: 754–768.

Austin, Peter C., Paul Grootendorst, Geoffrey M. Anderson 2007b: A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. Statistics in Medicine 26: 734–753.

Bach, K. 1980: Actions are not events. Mind 89: 114–120.

Bahadur, R.R., Leonard J. Savage 1956: The nonexistence of certain statistical procedures in nonparametric problems. Annals of Mathematical Statistics 27: 1115–1122. Also in: American Statistical Association, The Institute of Mathematical Statistics (ed) 1981: The Writings of Leonard Jimmie Savage — A Memorial Selection: 276–283.

Baker, Stuart G., Garrett M. Fitzmaurice, Laurence S. Freedman, Barnett S. Kramer 2006: Simple adjustment for randomized trials with nonrandomly missing or censored outcomes arising from informative covariates. Biostatistics 7: 29–40.

Baltagi, Badi H., Seuck Heun Song 2006: Unbalanced panel data: A survey. Statistical Papers 47: 493–523.

Bandeen-Roche, Karen, Kung-Yee Liang 2002: Modelling multivariate failure time association in the presence of competing risks. Biometrika 89: 299–314.

Bang, Heejung, James M. Robins 2005: Doubly robust estimation in missing data and causal inference models. Biometrics 61: 962–972.

Barber, Jennifer S., Susan A. Murphy, Natalya Verbitsky 2004: Adjusting for time-varying confounding in survival analysis. Sociological Methodology 34: 163–192.

Barndorff-Nielsen, O.E. 1978: Information and Exponential Families in Statistical Theory. New York: Wiley.

Barndorff-Nielsen, O.E., David R. Cox 1989: Asymptotic Techniques for Use in Statistics. London: Chapman & Hall.

Barndorff-Nielsen, O.E., David R. Cox 1994: Inference and Asymptotics. London: Chapman & Hall.

Barnes, Sunni A., Stacy R. Lindborg, John W. Seaman Jr. 2006: Multiple imputation techniques in small sample clinical trials. Statistics in Medicine 25: 233–245.

Barrios, Javier A. 2004: Generalized sample selection bias correction under RUM. Economics Letters 85: 129–132.

Bartholomew, David J. 1967: Stochastic Models for Social Processes. New York: Wiley.

Bayarri, María-Jesús, Morris H. DeGroot, Joseph B. Kadane 1987: What is the likelihood function? (with discussion). Pp. 3–27 in: S.S. Gupta, James O. Berger (eds): Statistical Decision Theory and Related Topics. Berlin: Springer.

Bayarri, María-Jesús, Morris H. DeGroot 1992: Difficulties and ambiguities in the definition of a likelihood function. Journal of the Italian Statistical Society 1: 1–15.

Bayne, Steven M. 2004: Kant on Causation. Albany: State University of New York Press.

Beaumont, Jean-François 2005: Calibrated imputation in surveys under a quasi-model-assissted approach. Journal of the Royal Statistical Society B 67: 445–458.

Bellio, R., E. Gori 2003: Impact evaluation of job training programmes: Selection bias in multilevel models. Journal of Applied Statistics 30: 893–907.

Belzil, Christian, Jörgen Hansen 2002: Unobserved ability and the return to schooling. Econometrica 70: 2075–2091.

Belzil, Christian, Jörgen Hansen 2007: A structural analysis of the correlated random coefficient wage regression model. Journal of Econometrics 140: 827–848.

Bembo, Oliver, Mark J. van der Laan 2007: A practical illustration of the importance of realistic individualized treatment rules in causal inference. Electronic Journal of Statistics 1: 574–596.

Bender, Stefan, J. Hilzendegen, Götz Rohwer, H. Rudolph 1996: Die IAB–Beschäftigtenstichprobe 1975–1990. Beiträge zur Arbeitsmarkt- und Berufsforschung 197. Nürnberg: Institut für Arbeitsmarkt- und Berufsforschung.

Bens, Arno 2006: Zur Auswertung haushaltsbezogener Merkmale mit dem ALLBUS 2004. ZA-Information 59: 143–156.

Berger, James O., Robert L. Wolpert 1988^2: The Likelihood Principle. Hayward: Institute of Mathematical Statistics.

Berger, Vance W. 2005a: The reverse propensity score to detect selection bias and correct for baseline imbalances. Statistics in Medicine 24: 2777–2787.

Berger, Vance W. 2005b: Quantifying the magnitude of baseline covariate imbalances resulting from selection bias in randomized clinical trials (with discussion). Biometrical Journal 47: 119–139.

Bergström, R., P.-A. Edin 1992: Time aggregation and the distributional shape of unemployment duration. Journal of Applied Econometrics 7: 5–30.

Berk, Richard A. 2004: Regression Analysis. A Constructive Critique. Thousand Oaks: Sage.

Berk, Richard A., Subash C. Ray 1982: Selection bias in sociological data. Social Science Research 11: 352–398.

Berman, Simeon M. 1963: Note on extreme values, competing risks and semi-Markov processes. Annals of Mathematical Statistics 34: 1104–1106.

Bernert, C. 1983: The career of causal analysis in American sociology. British Journal of Sociology 34: 230–254.

Betensky, Rebecca A. 2000: On nonidentifiability and noninformative censoring for current status data. Biometrika 87: 218–221.

Betensky, Rebecca A., D.M. Finkelstein 1999: A non-parametric maximum likelihood estimator for bivariate interval censored data. Statistics in Medicine 18: 3089–3100.

Beunckens, Caroline, Geert Molenberghs, Herbert Thijs, Geert Verbeke 2007: Incomplete hierarchical data. Statistical Methods in Medical Research 16: 457–492.

Bertoin, Jean 2006: Random Fragmentation and Coagulation Processes. Cambridge: Cambridge University Press.

Bickel, Peter J., Chris A.J. Klaassen, Ya'acov Ritov, Jon A. Wellner 1993: Efficient and Adaptive Estimation for Semiparametric Models. Berlin: Springer.

Bickel, Peter J., Jaimyoung Kwon 2001: Inference for semiparametric models: Some questions and an answer (with discussion). Statistica Sinica 11: 863–960.

Bickel, Peter J., Ya'acov Ritov 2003: Nonparametric estimators which can be "plugged-in". Annals of Statistics 31: 1033–1053.

Bilias, Yannis, Minggao Gu, Zhiliang Ying 1997: Towards a general asymptotic theory for Cox model with staggered entry. Annals of Statistics 25: 662–682.

Birmingham, Jolene, Andrea Rotnitzky, Garrett M. Fitzmaurice 2003: Pattern-mixture and selection models for analysing longitudinal data with monotone missing patterns. Journal of the Royal Statistical Society B 65: 275–297.

Blossfeld, Hans-Peter, Gerald Prein 1998: Rational Choice Theory and Large Scale Data Analysis. Boulder (CO): Westview.

Blossfeld, Hans-Peter, Götz Rohwer 2002^2: Techniques of Event History Modeling. New Approaches to Causal Analysis. Hillsdale: Erlbaum.

Blume, Jeffrey D., Li Su, Remigio M. Olveda, Stephen T. McGarvey 2007: Statistical evidence for GLM regression parameters: A robust likelihood approach. Statistics in Medicine 26: 2919–2936.

Blumer, Herbert 1956: Sociological analysis and the 'variable'. American Sociological Review 21: 683–690.

Blundell, Richard, Lorraine Dearden, Barbara Sianesi 2005: Evaluating the effect of education on earnings: Models, methods and results from the National Child Development Survey. Journal of the Royal Statistical Society A 168: 473–512.

Blundell, Richard, Amanda Gosling, Hidehiko Ichimura, Costas Meghir 2007: Changes in the distribution of male and female wages accounting for employment composition using bounds. Econometrica 75: 323–363.

Borgan, Ørnulf, Rosemeire L. Fiaccone, Robin Henderson, Mauricio L. Barreto 2007: Dynamic analysis of recurrent event data with missing observations, with application to infant diarrhoea in Brazil. Scandinavian Journal of Statistics 34: 53–69.

Boscardin, W. John, Xiaohong Yan, Weng Kee Wong 2007: A reanalysis of a longitudinal scleroderma clinical trial using non-ignorable missingness models. Journal of Statistical Planning and Inference 137: 3848–3858.

Bottai, Matteo 2003: Confidence regions when Fisher information is zero. Biometrika 90: 73–84.

Bowman, F. DuBois, Amita K. Manatunga 2005: A joint model for longitudinal data profiles and associated event risks with application to a depression study. Applied Statistics 54: 301–316.

Brady, Michael Emmett 1988: J.M. Keynes's position on the general applicability of mathematical, logical and statistical methods in economics and social science. Synthese 76: 1–24.

Brame, Robert, Raymond Paternoster 2003: Missing data problems in criminological research: Two case studies. Journal of Quantitative Criminology 19: 55–78.

Brand, Jennie E., Yu Xie 2007: Identification and estimation of causal effects with time-varying treatments and time-varying outcomes. Sociological Methodology 37: 393–434.

Bravo, José I., Íñigo de Fuentes, Arturo J. Fernández 2006: Survival estimation with missing censoring times for the generalized Koziol-Green model. Communications in Statistics–Theory and Methods 35: 363–372.

Bray, Bethany Cara, Daniel Almirall, Rick S. Zimmerman, Donald Lynam, Susan Murphy 2006: Assessing the total effect of time-varying predictors in prevention research. Prevention Science 7: 1–17.

Brémaud, Pierre 1981: Point Processes and Queues. Martingale Dynamics. Berlin: Springer.

Breslow, Norman E., Jon A. Wellner 2007: Weighted likelihood for semiparametric models and two-phase stratified samples, with applications to Cox regression. Scandinavian Journal of Statistics 34: 86–102.

Bretagnolle, J., C.J. Huber-Carol 1985 : Sous-estimation des contrastes due à l'oublie de variables pertinentes dans le modèle de Cox pour les durées de survie avec censure. Comptes Rendus de l'Académie des Sciences Paris, Série I, 300 : 359–362.

Bretagnolle, J., C.J. Huber-Carol 1988: Effects of omitting covariates in Cox's model for survival data. Scandinavian Journal of Statistics 15: 125–138.

Brillinger, David R. 1983: A generalized linear model with "Gaussian" regressor variables. Pp. 97–114 in: Bickel, Peter J., Kjell A. Doksum, J.L. Hodges (eds) 1983: A Festschrift for Erich L. Lehmann. Belmont, CA: Wadsworth.

Brillinger, David R. 1986: The natural variability of vital rates (with discussion). Biometrics 42: 693–734.

Brookhart, M. Alan, Mark J. van der Laan 2006: A semiparametric model selection criterion with applications to the marginal structural model. Computational Statistics & Data Analysis 50: 475–498.

Brookhart, M. Alan, Sebastian Schneeweiss, Kenneth J. Rothman, Robert J. Glynn, Jerry Avorn, Til Stürmer 2006: Variable selection for propensity score models. American Journal of Epidemiology 163: 1149–1156.

Brown, B.M., You-Gan Wang 2007: Induced smoothing for rank regression with censored survival times. Statistics in Medicine 26: 828-836.

Brown, Lawrence D., T. Tony Cai, Anirban DasGupta 2001: Interval estimation for a binomial proportion (with discussion). Statistical Science 16: 101–133.

Brown, Lawrence D., Anirban DasGupta, L.R. Haff, W.E. Strawderman 2006: The heat equation and Stein's identity: Connections, applications. Journal of Statistical Planning and Inference 136: 2254–2278.

Brumback, Babette A., Miguel A. Hernán, Sebastien J.P.A. Haneuse, James M. Robins 2004: Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. Statistics in Medicine 23: 749–767.

Bryan, Jenny, Zhuo Yu, Mark J. van der Laan 2004: Analysis of longitudinal marginal structural models. Biostatistics 5: 361–380.

Bryson, Maurice C. 1976: The Literary Digest poll: Making of a statistical myth. American Statistician 30: 184–185.

Brüderl, Josef, Andreas Diekmann, Henriette Engelhardt 1997: Erhöht eine Probeehe das Scheidungsrisiko? Eine empirische Untersuchung mit dem Familiensurvey. Kölner Zeitschrift für Soziologie und Sozialpsychologie 49: 205–222.

Buckley, J., I. James 1979: Linear regression with censored data. Biometrika 66: 429–436.

Bundesregierung, 2001: Lebenslagen in Deutschland. Daten und Fakten. Materialband zum ersten Armuts- und Reichtumsbericht der Bundesregierung. BundestagsdrucksachenNr. 14:/ 5990

Bunge, Mario 1963: Causality. Cleveland: Meridian Books.

Burkam, David T., Valerie E. Lee 1998: Effects of monotone and nonmonotone attrition on parameter estimates in regression models with educational data. Demographic effects on achievement, aspiration, and attitudes. The Journal of Human Resources 33: 555–574.

Burton, Jonathan, Heather Laurie, Peter Lynn 2006: The long-term effectiveness of refusal conversion procedures on longitudinal surveys. Journal of the Royal Statistical Society A 169: 459–478.

Cacoullos, T., V. Papathanasiou 1992: Lower variance bounds and a new proof of the central limit theorem. Journal of Multivariate Analysis 43: 173–184.

Cahalan, Don 1989: The Digest poll rides again. Public Opinion Quarterly 53: 129–133.

Cai, J., Ross L. Prentice 1995: Estimating equations for hazard ratio parameters based on correlated failure time data. Biometrika 82: 151–164.

Cai, Jianwen, Jianqing Fan, Jiancheng Jiang, Haibo Zhou 2007: Partially linear hazard regression for multivariate survival data. Journal of the American Statistical Association 102: 538–551.

Cai, Jianwen, Jianqing Fan, Jiancheng Jiang, Haibo Zhou 2008: Partially linear hazard regression with varying coefficients for multivariatesurvival data. Journal of the Royal Statistical Society B 70: 141–158.

Cai, Jianwen, Jinheum Kim 2003: Nonparametric quantile estimation with correlated failure time data. Lifetime Data Analysis 9: 357–371.

Caliendo, Marco, Reinhard Hujer 2006: The microeconometric estimation of treatment effects – An overview. Allgemeines Statistisches Archiv 90: 199–215.

Campbell, G. 1981a: Nonparametric bivariate estimation with randomly censored data. Biometrika 68: 417–422.

Campbell, G. 1981b: Asymptotic properties of several nonparametric multivariate distribution function estimators under random censorship. Pp. 243–256 in: John Crowley, R.A. Johnson (eds) 1981: Survival Analysis. Columbus: Institute of Mathematical Statistics Lecture Notes, Monograph Series, 2.

Cantoni, Eva, Xavier de Luna 2006: Non-parametric adjustment for covariates when estimating a treatment effect. Nonparametric Statistics 18: 227–244.

Card, David 2001: Estimating the return to schooling: Progress on some persistent econometric problems. Econometrica 69: 1127–1160.

Carpenter, James R., Michael G. Kenward, Stijn Vansteelandt 2006: A comparison of multiple imputation and doubly robust estimation for analyses with missing data. Journal of the Royal Statistical Society A 169: 571–584.

Carpenter, James R., Michael G. Kenward, Ian R. White 2007: Sensitivity analysis after multiple imputation under missing at random: A weighting approach. Statistical Methods in Medical Research 16: 259–275.

Cartwright, Nancy 1989: Nature's Capacities and their Measurement. Oxford: Clarendon Press.

Cartwright, Nancy 1999: The Dappled World. A Study of the Boundaries of Science. Cambridge: Cambridge University Press.

Cartwright, Nancy 2002: Against modularity, the causal Markov condition, and any link between the two: Comments on Hausman and Woodward. British Journal for the Philosophy of Science 53: 411–453.

Cartwright, Nancy 2006: From metaphysics to method: Comments on manipulability and the causal Markov condition. British Journal for the Philosophy of Science 57: 197–218.

Castro, Jordi 2006: Minimum-distance controlled pertubation methods for large-scale tabular data protection. European Journal of Operational Research 171: 39–52.

Catchpole, E.A., B.J.T. Morgan, G. Tavecchia 2008: A new method for analysing discrete life history data with missing covariate values. Journal of the Royal Statistical Society B 70: 445–460.

Cator, Eric A. 2004: On the testability of the CAR assumption. Annals of Statistics 32: 1957–1980.

Causey, Beverley D., Lawrence H. Cox, Lawrence R. Ernst 1985: Applications of transportation theory to statistical problems. Journal of the American Statistical Association 80: 903–909.

Celeux, Gilles, F. Forbes, C.P. Robert, D.M. Titterington 2006: Deviance information criteria for missing data models (with discussion). Bayesian Analysis 1: 651–706.

Celeux, Gilles, Jean-Michel Marin, Christian P. Robert 2006: Iterated importance sampling in missing data problems. Computational Statistics & Data Analysis 50: 3386–3404.

Chambers, E.A., David R. Cox, D.R. 1967: Discrimination between alternative binary response models. Biometrika 54: 573–578.

Chang, Shu-Hui, Shinn-Jia Tzeng 2006: Nonparametric estimation of sojourn time distributions for truncated serial event data–a weight-adjusted approach. Lifetime Data Analysis 12: 53–67.

Chateauneuf, Alain, Jean-Yves Jaffray 1989: Some characterizations of lower probabilities and other monotone capacities through the use of Moebius inversion. Mathematical Social Sciences 17: 263–283.

Chatterjee, Nilanjan, Joanna Shih 2001: A bivariate cure-mixture approach for modeling familial association in diseases. Biometrics 57: 779–786.

Chavance, M. 2004: Handling missing items in quality of life studies. Communications in Statistics–Theory and Methods 33: 1371–1383.

Chen, Bingshu E., Joan L. Kramer, Mark H. Greene, Philip S. Rosenberg 2008: Competing risks analysis for correlated failure time data. Biometrics 64: 172–179.

Chen, Hua, Zhi Geng, Jinzhu Jia 2007: Criteria for surrogate end points. Journal of the Royal Statistical Society B 69: 919–932.

Chen, Hua Yun 2002: Double-semiparametric method for missing covariates in Cox regression models. Journal of the American Statistical Association 97: 565–576.

Chen, Hua Yun, Roderick J. Little 1999: Proportional hazards regression with missing covariates. Journal of the American Statistical Association 94: 896–908.

Chen, Hua Yun, Roderick J. Little 2001: A profile conditional likelihood approach for the semiparametric transformation regression model with missing covariates. Lifetime Data Analysis 7: 207–224.

Chen, J., J.N.K. Rao, R.R. Sitter 2000: Efficient random imputation for missing data in complex surveys. Statistica Sinica 10: 1153–1169.

Chen, Jianwei, Jianqin Fan, Kim-Hung Li, Haibo Zhou 2006: Local quasi-likelihood estimation with data missing at random. Statistica Sinica 16: 1071–1100.

Chen, Jinbo, Norman E. Breslow 2004: Semiparametric efficient estimation for the auxiliary outcome problem with the conditional mean model. Canadian Journal of Statistics 32: 359–372.

Chen, Kani 2001: Generalized case-cohort sampling. Journal of the Royal Statistical Society B 63: 791–809.

Chen, Kani, Jia Shen, Zhiliang Ying 2005: Rank estimation in partial linear model with censored data. Statistica Sinica 15: 767–779.

Chen, Qiangxia, Joseph G. Ibrahim 2006: Semiparametric models for missing covariate and response data in regression models. Biometrics 62: 177–184.

Chen, Qiangxia, Donglin Zeng, Joseph G. Ibrahim 2007: Sieve maximum likelihood estimation for regression models with covariates missing at random. Journal of the American Statistical Association 102: 1309–1317.

Chen, Song Xi, Jing Qin 2006: An empirical likelihood method in mixture models with incomplete classifications. Statistica Sinica 16: 1101–1115.

Chen, Y.Q., N.P. Jewell, X. Lei, S.C. Cheng 2005: Semiparametric estimation of proportional mean residual life model in presence of censoring. Biometrics 61: 170–178.

Chen, Yinzhong, Jun Shao 1999: Inference with survey data imputed by hot deck when imputed values are nonidentifiable. Statistica Sinica 9: 361–384.

Chen, Yuguo, Persi Diaconis, Susan P. Holmes, Jun S. Liu 2005: Sequential Monte Carlo methods for statistical analysis of tables. Journal of the American Statistical Association 100: 109–120.

Chen, Yuguo, Ian H. Dinwoodie, Seth Sullivant 2006: Sequential importance sampling for multiway tables. Annals of Statistics 34: 523–545.

Cheng, Bin, Jun Shao, Bob Zhong 2005: Last observation analysis in ANOVA and ANCOVA. Statistica Sinica 15: 857–870.

Chernozhukov, Victor, Christian Hansen 2005: An IV model of quantile treatment effects. Econometrica 73: 245–261.

Chernozhukov, Victor, Christian Hansen 2006: Instrumental quantile regression inference for structural and treatment effect models. Journal of Econometrics 132: 491–525.

Chernozhukov, Victor, Han Hong, Elie Tamer 2007: Estimation and confidence regions for parameter sets in econometric models. Econometrica 75: 1243–1284.

Cheung, Ying Kuen 2005: Exact two-sample inference with missing data. Biometrics 61: 524–531.

Chib, Siddartha 2007: Analysis of treatment response data without the joint distribution of potential outcomes. Journal of Econometrics 140: 401–412.

Chib, Siddartha, Liana Jacobi 2007: Modeling and calculating the effect of treatment at baseline from panel outcomes. Journal of Econometrics 140: 781–801.

Chung, C.-F., A.S. Goldberger 1984: Proportional projections in limited dependent variable models. Econometrica 52: 531–534.

Clarke, Paul S., Peter W.F. Smith 2004: Interval estimation for log-linear models with one variable subject to non-ignorable non-response. Journal of the Royal Statistical Society B 66: 357–368.

Clarke, Paul S., Peter W.F. Smith 2005: On maximum likelihood estimation for log-linear models with non-ignorable non-response. Statistics & Probability Leters 73: 441–448.

Clogg, C.C. 1992: The impact of sociological methodology on statistical methodology (with discussion). Statistical Science 7: 183–207.

Clogg, C.C., A. Haritou 1997: The regression method of causal inference and a dilemma confronting this method. Pp. 83–112 in: McKim, V.R., S.P. Turner (eds) 1997: Causality in Crisis? Statistical Methods and the Search for Causal Knowledge in the Social Sciences. University of Notre Dame Press: Notre Dame.

Cochran, William G. 1977^3: Sampling Techniques. New York: Wiley.

Cole, Bernard F., Marco Bonetti, Alan M. Zaslavsky, Richard D. Gelber 2005: Multistate Markov chain model for longitudinal, categorical quality-of-life data subject to non-ignorable missingness. Statistics in Medicine 24: 2317–2334.

Commenges, Daniel, Anne Gégout-Petit 2005: Likelihood inference for incompletely observed stochastic processes: Ignorability conditions. arXiv:math.ST/0507151.

Commenges, Daniel, Anne Gégout-Petit 2007: Likelihood for generally coarsened observations from multistate or counting process models. Scandinavian Journal of Statistics 34: 432–450.

Commenges, Daniel, Pierre Joly, Anne Gégout-Petit, Benoit Liquet 2007: Choice between semi-parametric estimators of Markov and non-Markov multi-state models from coarsened observations. Scandinavian Journal of Statistics 34: 33–52.

Commenges, Daniel, Virginie Rondeau 2006: Relationships between derivatives of the observed and full loglikelihoods and application to Newton-Raphson algorithm. The International Journal of Biostatistics 2: Article 4.

Cong, Xiuyu J., Guosheng Yin, Yu Shen 2007: Marginal analysis of correlated failure time data with informative cluster size. Biometrics 63: 663–672.

Conniffe, Denis, Vanessa Gash, Philip J. O´Connell 2000: Evaluating state programs: "Natural Experiments" and propensity scores. The Economic and Social Review 31: 283–308.

Constant, Amelie, Douglas S. Massey 2003: Self-selection, earnings, and outmigration: A longitudinal study of immigrants to Germany. Journal of Population Economics 16: 631–653.

Cook, Richard J., Leilei Zeng, Grace Y. Yi 2004: Marginal analysis of incomplete longitudinal binary data: A cautionary note on LOCF imputation. Biometrics 60: 820–828.

Cook, Thomas D. 2008: "Waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics and economics. Journal of Econometrics 142: 636–654.

Copas, John B., H.G. Li 1997: Inference for non-random samples (with discussion). Journal of the Royal Statistical Society B 59: 55–95.

Copas, John B., Shinto Eguchi 2001: Local sensitivity for selectivity bias. Journal of the Royal Statistical Society B 63: 871–895.

Copas, John B., Shinto Eguchi 2005: Local model uncertainty and incomplete-data bias (with discussion). Journal of the Royal Statistical Society B 67: 459–513.

Cox, David R. 1990: Role of models in statistical analysis. Statistical Science 5: 169–174.

Cox, David R. 1992: Causality: Some statistical aspects. Journal of the Royal Statistical Society A 155: 291–301.

Cox, David R. 2007: On a generalization of a result of W.G. Cochran. Biometrika 94: 755–759.

Cox, David R., David Oakes 1984: Analysis of Survival Data. London: Chapman & Hall.

Cox, David R., Nancy Reid 1987: Parameter orthogonality and approximate conditional inference. Journal of the Royal Statistical Society B 49: 1–39.

Cox, David R., Nanny Wermuth 2004: Causality: A statistical view. International Statistical Review 72: 285–305.

Cox, David, John Little, Donald O'Shea 2002^2: Ideals, Varieties, and Algorithms. Berlin: Springer.

Craiu, Radu V., Benjamin Reiser 2006: Inference for the dependent competing risks model with masked causes of failure. Lifetime Data Analysis 12: 21–33.

Cross, Philip J., Charles F. Manski 2002: Regression, short and long. Econometrica 70: 357–368.

Crouchley, Rob, Mojtaba Ganjali 2002: The common structure of several models for non-ignorable dropout. Statistical Modeling 2: 39–62.

Crowder, Martin 1996: On assessing independence of competing risks when failure times are discrete. Lifetime Data Analysis 2: 195–209.

Crowder, Martin 1997: A test for independence of competing risks with discrete failure times. Lifetime Data Analysis 3: 215–223.

Crowder, Martin 2000: Characterizations of competing risks in terms of independent-risks proxy models. Scandinavian Journal of Statistics 27: 57–64.

Crowder, Martin 2001: Corrected *p*-values for tests based on estimated nuisance parameters. Statistics and Computing 11: 359–365.

Currie, I.D. 1996: A note on Buckley–James estimators for censored data. Biometrika 83: 912–915.

Cuzick, Jack, Peter Sasieni, Jonathan Myles, Jonathan Tyrer 2007: Estimating the effect of treatment in a proportional hazards model in the presence of non-compliance and contamination. Journal of the Royal Statistical Society B 69: 565–588.

Dabrowska, Dorota M. 1988: Kaplan–Meier estimate on the plane. Annals of Statistics 16: 1475–1489.

Dabrowska, Dorota M. 1989: Kaplan–Meier estimate on the plane: Weak convergence, LIL, and the bootstrap. Journal of Multivariate Analysis 29: 308–325.

Dabrowska, Dorota M. 2005: Quantile regression in transformation models. Sankhyā 67: 153–186.

Dabrowska, Dorota M., W. Lee 1996: Nonparametric estimation of transition probabilities in a two-stage duration model. Nonparametric Statistics 7: 75–103.

D'Agostino, Ralph B., M.L. Lee, A.J. Belanger, L.A. Cupples, K. Anderson, W.B. Kannel, W.B. 1990: Relation of pooled logistic regression to time dependent Cox regression analysis: The Framingham Heart Study. Statistics in Medicine 9: 1501–1515.

D'Agostino, Ralph B., Donald B. Rubin 2000: Estimating and using propensity scores with partially missing data. Journal of the American Statistical Association 95: 749–759.

D'Agostino, Ralph B. 2004: Propensity score estimation with missing data. Pp. 163–174 in: Gelman, Andrew, Xiao-Li Meng (eds) 2004: Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives. An Essential Journey with Donald Rubin's Statistical Family. New York: Wiley.

Dahl, Gordon B. 2002: Mobility and the return to education: Testing a Roy model with multiple markets. Econometrica 70: 2367–2420.

Dahlhaus, Rainer, Michael Eichler 2003: Causality and graphical models in time series analysis (with discussion). Pp. 115–144 in: Peter J. Green, Nils Lid Hjort, Sylvia Richardson (eds.) 2003: Highly Structured Stochastic Systems. Oxford: Oxford University Press.

Daley, D.J., S. Vere-Jones 1988: An Introduction to the Theory of Point Processes. Berlin: Springer.

D'Angelo, Gina, Lisa Weissfeld 2007: Covariates missing by design: Comparison of the efficient score to other weighted methods. Statistics in Medicine: 2137–2153.

Dantsin, Evgeny, Vladik Kreinovich, Alexander Wolpert, Gang Xiang 2006: Population variance under interval uncertainty: A new algorithm. Reliable Computing 12: 273–280.

Das, Mitali 2004: Simple estimators for nonparametric panel data models with sample attrition. Journal of Econometrics 120: 159–180.

Das, Mitali, Whitney K. Newey, Francis Vella 2003: Nonparametric estimation of sample selection models. Review of Economic Studies 70: 33–58.

Da Silva, Damiaõ N., Jean D. Opsomer 2006: A kernel smoothing method of adjusting for unit non-response in sample surveys. Canadian Journal of Statistics 34: 563–579.

Datta, Susmita, Jennifer Le-Rademacher, Somnath Datta 2007: Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and LASSO. Biometrics 63: 259–271.

Dauxois, Jean-Yves, Agathe Guilloux 2008: Nonparametric inference under competing risks and selection-biased sampling. Journal of Multivariate Analysis 99: 589–605.

Dauxois, Jean-Yves, Agathe Guilloux, Syed N.U.A. Kirmani 2005: Nonparametric estimation from proportional hazards competing risks data under selection bias. Nonparametric Statistics 17: 717–731.

Davidian, Marie, Anastasios A. Tsiatis, Selene Leon 2005: Semiparametric estimation of treatment effect in a pretest-posttest study with missing data (with discussion). Statistical Science 20: 261–301.

Davidson, Donald 1980: Essays on Actions and Events. Oxford: Oxford University Press.

Davies, Laurie 1995: Data features. Statistica Neerlandica 49: 185–245.

Davies, Laurie, Arne Kovac 2004: Densities, spectral densities and modality. Annals of Statistics 32: 1093–1136.

Davis, George C. 2000: A semantic interpretation of Haavelmo's structure of econometrics. Economics and Philosophy 16: 205–228.

Davis, George C. 2005: Clarifying the 'puzzle' between the testbook and the LSE approaches to econometrics: A comment on Cook's Kuhnian perspective on econometric modelling. Journal of Economic Methodology 12: 93–115.

Davis, W.A. 1988: Probabilistic theories of causation. Pp. 133–160 in: Fetzer, J.H. (ed) 1988: Probability and Causality. Essays in Honor of Wesley C. Salmon. D. Reidel: Dordrecht.

Davison, Anthony C., S. Sardy 2007: Resampling variance estimation in surveys with missing data. Journal of Official Statistics 23: 371–386.

Dawid, A. Philip 1979a: Conditional independence in statistical theory. Journal of the Royal Statistical Society B 41: 1–31.

Dawid, A. Philip 1979b: Some misleading arguments involving conditional independence. Journal of the Royal Statistical Society B 41: 249–252.

Dawid, A. Philip 1980: Conditional independence for statistical operations. Annals of Statistics 8: 598-617.

Dawid, A. Philip 2000: Causal inference without counterfactuals (with discussion). Journal of the American Statistical Association 95: 407–448.

Dawid, A. Philip 2002: Influence diagrams for causal modelling and inference. International Statistical Review 70: 161–189. Corrigendum: International Statistical Review (2002) 70: 437.

Dawid, A. Philip 2003: Causal inference using influence diagrams: The problem of partial compliance (with discussion). Pp. 45–81 in: Peter J. Green, Nils Lid Hjort, Sylvia Richardson (eds.) 2003: Highly Structured Stochastic Systems. Oxford: Oxford University Press.

Dawid, A. Philip 2004: Probability, causality and the empirical world: A Bayes–deFinetti–Popper–Borel synthesis. Statistical Science 19: 44–57.

Dawid, A. Philip 2006: Counterfactuals, hypotheticals and potential responses: A philosophical examination of statistical causality. Research Report 269. Department of Statistical Science, University College London.

Dawid, A. Philip, Julia Mortera 1989: Forensic identification with imperfect evidence. Biometrika 85: 835–849.

Deb, Partha, Pravin K. Trivedi 2006: Specification and simulated likelihood estimation of a non-normal treatment-outcome model with selection: Application to health care utilization. Econometrics Journal 9: 307–331.

de Cooman, Gert, Marco Zaffalon 2003: Updating with incomplete observations. Pp. 142–150 in: C. Meek, U. Kjærulff (eds) 2003: Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann.

de Cooman, Gert, Marco Zaffalon 2004: Updating beliefs with incomplete observations. Artificial Intelligence 159: 75–125.

Dehejia, Rajeev H. 2005: Practical propensity score matching: A reply to Smith and Todd. Journal of Econometrics 125: 355–364.

Dehejia, Rajeev H., Sadek Wahba 1999: Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. Journal of the American Statistical Association 94: 1053–1062.

Dehejia, Rajeev H., Sadek Wahba 2002: Propensity score-matching methods for nonexperimental causal studies. The Review of Economics and Statistics 84: 151–161.

Deistler, Manfred, Hans-Günther Seifert 1978: Identifiability and consistent estimability in econometric models. Econometrica 46: 969–980.

Del Bianco, P., R. Borgoni 2006: Handling dropout and clustering in longitudinal multicentre clinical trials. Statistical Modelling 6: 141–157.

Delecroix, Michel, Olivier Lopez, Valentin Patile 2008: Nonlinear censored regression using synthetic data. Scandinavian Journal of Statistics 35: 248–265.

De Loera, Jesús A. 2005: The many aspects of counting lattice points in polytopes. Mathematische Semesterberichte 52: 175–195.

de Luna, Xavier, Per Johansson 2006: Exogeneity in structural equation models. Journal of Econometrics 132: 527–543.

Demirtas, Hakan 2004: Simulation driven inferences for multiply imputed longitudinal datasets. Statistica Neerlandica 58: 466–482.

Demirtas, Hakan 2005: Multiple imputation under Bayesianly smoothed pattern-mixture models for non-ignorable drop-out. Statistics in Medicine 24: 2345–2363.

Demirtas, Hakan, Donald Hedeker 2007: Gaussianization-based quasi-imputation and expansion strategies for incomplete correlated binary responses. Statistics in Medicine 26: 782–799.

Demisse, Serkalem, Michael P. LaValley, Nicholas J. Horton, Robert J. Glynn, L. Adrienne Cupples 2003: Bias due to missing exposure data using complete-case analysis in the proportional hazards regression model. Statistics in Medicine 22: 545–557.

Dempster, Athur P. 1990: Causality and statistics. Journal of Statistical Planning and Inference 25: 261–278.

Desai, Tejas A., Pranab K. Sen 2006: Information attainable in some randomly incomplete data models. Journal of Statistical Planning and Inference 136: 2309–2326.

Devroye, Luc, Lázló Győrfi 1990: No empirical measure can converge in the total variation sense for all distributions. Annals of Statistics 18: 1496–1499.

Diaconis, Persi, Bradley Efron 1985: Testing for independence in a two-way table: New interpretations of the chi-square statistic (with discussion). Annals of Statistics 13: 845–913.

Didelez, Vanessa 2002: ML- and semiparametric estimation in logistic models with incomplete covariate data. Statistica Neerlandica 56: 330–345.

Didelez, Vanessa 2007: Graphical models for composable finite Markov processes. Scandinavian Journal of Statistics 34: 169–185.

Didelez, Vanessa 2008: Graphical models for marked point processes based on local independence. Journal of the Royal Statistical Society B 70: 245–264.

Didelez, Vanessa, Nuala Sheehan 2005: Mendelian randomization and instrumental variables: What can and what can't be done. Technical Report 05-02, Department of Health Sciences, University of Leicester.

DiPrete, Thomas A., Markus Gangl 2004: Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. Sociological Methodology 34: 271–310.

Diekmann, Andreas, David Wyder 2002: Vertrauen und Reputationseffekte bei Internet-Auktionen. Kölner Zeitschrift für Soziologie und Sozialpsychologie 54: 674–693.

Diggle, Peter, Daniel Farewell, Robin Henderson 2007: Analysis of longitudinal data with drop-out: Objectives, assumptions and a proposal (with discussion). Applied Statistics 56: 499–550.

Diggle, Peter, Michael G. Kenward 1994: Informative drop-out in longitudinal data analysis (with discussion). Applied Statistics 43: 49–93.

Dignam, James J., Kelly Wieand, Paul J. Rathouz 2007: A missing data approach to semi-competing risks problems. Statistics in Medicine 26: 837–856.

Ding, Aidong Adam, Weijing Wang 2007: Inference for bivariate survival data by copula models adjusted for the boundary effect. Communications in Statistics – Theory and Methods 36: 2927–2936.

Dobra, Adrian 2003: Markov bases for decomposable graphical models. Bernoulli 9: 1093–1108.

Dobra, Adrian, Stephen E. Fienberg 2000: Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. Proceedings of the National Academy of Sciences 97: 11885–11892.

Dobra, Adrian, Claudia Tebaldi, Mike West 2006: Data augmentation in multiway contingency tables with fixed marginal totals. Journal of Statistical Planning and Inference 136: 355–372.

Dobson, Angela, Robin Henderson 2003: Diagnostics for joint longitudinal and dropout time modeling. Biometrics 59: 741–751.

Doksum, Kjell A., M. Gasko 1990: On a correspondence between models in binary regression analysis and in survival analysis. International Statistical Review 58: 243–252.

Donnet, Sophie, Adeline Samson 2007: Estimation of parameters in incomplete data models defined by dynamical systems. Journal of Statistical Planning and Inference 137: 2815–2831.

Dowe, Phil 2000: Physical Causation. Cambridge: Cambridge University Press.

Drton, Mathias, Steen A. Andersson, Michael D. Perlman 2006: Conditional independence models for seemingly unrelated regressions with incomplete data. Journal of Multivariate Analysis 97: 385–411.

Du, C., D. Kurowicka, R.M. Cooke 2006: Techniques for generic probabilistic inversion. Computational Statistics & Data Analysis 50: 1164–1187.

Dubra, Juan, Federico Echenique 2004: Information is not about measurability. Mathematical Social Sciences 47: 177–185.

Dudley, Richard M. 2002: Real Analysis and Probability. Cambridge: Cambridge University Press.

Dufouil, Carole, Carol Brayne, David Clayton 2004: Analysis of longitudinal studies with death and drop-out: a case study. Statistics in Medicine 23: 2215–2226.

Dupačová, Jitka, Roger Wets 1988: Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems. Annals of Statistics 16: 1517–1549.

Dupuy, Jean-François, Mounir Mesbah 2004: Estimation of the asymptotic variance of semiparametric maximum likelihood estimators in the Cox model with missing time-dependent covariate. Communications in Statistics – Theory and Methods 33: 1385–1401.

Dupuy, Jean-François, Ion Grama, Mounir Mesbah 2006: Asymptotic theory for the Cox model with missing time-dependent covariate. Annals of Statistics 34: 903–924.

Durant, Gabriele B., Chris Skinner 2006: Using data augmentation to correct for non-ignorable non-response when surrogate data are available: An application to the distribution of hourly pay. Journal of the Royal Statistical Society A 169: 605–623.

Dyer, Martin, Ravi Kannan, John Mount 1997: Sampling contingency tables. Random Structures and Algorithms 10: 487–506.

Ebrahimi, Nader, Daniel Molefe, Zhiliang Ying 2003: Identifiability and censored data. Biometrika 90: 724–727.

Eells, Ellery 1991: Probabilistic Causality. Cambridge: Cambridge University Press.

Efromovich, Sam 2004a: Distribution estimation for biased data. Journal of Statistical Planning and Inference 124: 1–43.

Efromovich, Sam 2004b: Density estimation for biased data. Annals of Statistics 32: 1137–1161.

Efron, Bradley 1967:The two sample problem with censored data. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 4: 831–883.

Efron, Bradley 1988: Logistic regression, survival analysis, and the Kaplan–Meier curve. Journal of the American Statistical Association 83: 414–425.

Efron, Bradley 1994: Missing data, imputation, and the bootstrap (with discussion). Journal of the American Statistical Association 89: 463–479.

Egleston, Brian L., Daniel O. Scharfstein, Ellen E. Freeman, Sheila K. West 2007: Causal inference for non-mortality outcomes in the presence of death. Biostatistics 8: 526–545.

Elliott, Michael R., Nicolas Stettler 2007: Using a mixture model for multiple imputation in the presence of outliers: The 'Healthy for life' project. Applied Statistics 56: 63–78.

Eltinge, John L. 2004: Nonresponse adjustment in government statistical agencies: Constraints, inferential goals, and robustness issues. Pp. 111–115 in: Gelman, Andrew, Xiao-Li Meng (eds) 2004: Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives. An Essential Journey with Donald Rubin's Statistical Family. New York: Wiley.

Engelhardt, Henriette 1999: Lineare Regression mit Selektion: Möglichkeiten und Grenzen der Heckman-Korrektur. Kölner Zeitschrift für Soziologie und Sozialpsychologie 51: 706–723.

Engle, Robert F., David F. Hendry, Jean-François Richard 1983: Exogeneity. Econometrica 51: 277–304.

Epstein, Lee, Daniel E. Ho, Gary King, Jeffrey A. Segal 2005: The Supreme Court during crisis: How war affects only non-war cases. New York University Law Review 80 (1): 1–116.

Epstein, R.J. 1987: A History of Econometrics. Amsterdam: North-Holland.

Estevao, Victor M., Carl-Erik Särndal 2004: Borrowing strength is not the best technique within a wide class of design-consistent domain estimators. Journal of Official Statistics 20: 645–669.

Fahrmeir, Ludwig 1990: Maximum likelihood estimation in misspecified generalized linear models. Statistics 21: 487–502.

Fan, Juanjuan, Ross L. Prentice 2002: Covariate-adjusted dependence estimation on a finite bivariate failure time region. Statistica Sinica 12: 689–705.

Fang, Yixin, Lincheng Zhao 2006: Approximation to the distribution of LAD estimators for censored regression by random weighting method. Journal of Statistical Planning and Inference 136: 1302–1316.

Favre, Anne-Catherine, Alina Matei, Yves Tillé 2005: Calibrated random imputation for qualitative data. Journal of Statistical Planning and Inference 128: 411–425.

Fay, Robert E. 1996: Alternative paradigms for the analysis of imputed survey data. Journal of the American Statistical Association 91: 490–498.

Feng, De-Jun, Ding Feng 2004: On a statistical framework for estimation from random set observations. Journal of Theoretical Probability 17: 85–110.

Fernández, Arturo J. 2006: Bounding maximum likelihood estimates based on incomplete ordered data. Computational Statistics & Data Analysis 50: 2014–2027.

Fernholz, Luisa Turrin 1983: von Mises Calculus for Statistical Functionals. Berlin: Springer.

Ferson, Scott, Lev Ginzburg, Vladik Kreinovich, Luc Longpré, Monica Aviles 2002: Computing variance for interval data is NP-hard. ACM SIGACT News 33: 108–118.

Fine, Jason P., Jun Yan, Michael R. Kosorok 2004: Temporal process regression. Biometrika 91: 683–703.

Firpo, Sergio 2007: Efficient semiparametric estimation of quantile treatment effects. Econometrica 75: 259–276.

Fisher, Ronald A. 1922: On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society A 222: 309–368.

Fisher, Ronald A. 1925: Theory of statistical estimation. Proceedings of the Cambridge Philosophical Society 22: 700–725.

Fishman, George S., David S. Rubin 1998: Best- and worst-case variances when bounds are available for the distribution function. Computational Statistics & Data Analysis 29: 35–53.

Fitzgerald, John, Peter Gottschalk, Robert Moffitt 1998: An analysis of sample attrition in panel data. The Michigan Panel Study of Income Dynamics. The Journal of Human Resources 33: 251–299.

Fitzmaurice, Garrett M. 2003: Methods for handling dropouts in longitudinal clinical trials. Statistica Neerlandica 57: 75–99.

Fitzmaurice, Garrett M., Nan M. Laird, Lucy Shneyer 2001: An alternative parametrization of the general linear mixture model for longitudinal data with non-ignorable drop-outs. Statistics in Medicine 20: 1009–1021.

Fitzmaurice, Garrett M., Stuart R. Lipsitz, Geert Molenberghs, Joseph G. Ibrahim 2005: A protective estimator for longitudinal binary data subject to non-ignorable non-monotone missingness. Journal of the Royal Statistical Society A 168: 723–735.

Fitzmaurice, Garrett M., Stuart R. Lipsitz, Joseph G. Ibrahim, Richard Gelber, Steven Lipshultz 2006: Estimation in regression models for longitudinal binary data with outcome-dependent follow-up. Biostatistics 7: 469–485.

Fleming, Thomas R., David P. Harrington 1991: Counting Processes and Survival Analysis. New York: Wiley.

Florens, J.P., M. Mouchart 1982: A note on noncausality. Econometrica 50: 583–589.

Florens, J.P., M. Mouchart 1985: A linear theory for noncausality. Econometrica 53: 157–175.

Forster, Jonathan J., Peter W.F. Smith 1998: Model based inference for categorical survey data subject to non-ignorable non-response. Journal of the Royal Statistical Society B 60: 57–70.

Frangakis, Constantine E., Donald B. Rubin 2002: Principal stratification in causal inference. Biometrics58: 21–29.

Frangakis, Constantine E., Donald B. Rubin, Ming-Wen An, Ellen MacKenzie 2007: Principal stratification designs to estimate input data missing due to death (with discussion). Biometrics 63: 641–662.

Frank, Kenneth, Kyung-Seok Min 2007: Indices of robustness for sample representation. Sociological Methodology 37: 349–392.

Freedman, David A. 2005: Statistical Models: Theory and Practice. Cambridge: Cambridge University Press.

Frigg, Roman, Stephan Hartmann 2005: Scientific Models. Pp. 740–749 in: Sahotar Sarkar et al. (eds): The Philosophy of Science: An Encyclopedia, vol. II. New York: Routledge.

Frick, Joachim R., Markus M. Grabka 2003: Missing data in the German SOEP: Incidence, imputation and its impact on the income distribution. Discussion Papers 376: Berlin: DIW.

Frölich, Markus 2004: A note on the role of the propensity score for estimating average treatment effects. Econometric Reviews 23: 167–174.

Frölich, Markus 2005: Matching estimators and optimal bandwidth choice. Statistics and Computing 15: 197–215.

Frölich, Markus 2006: Semiparametric estimation of conditional mean functions with missing data. Empirical Economics 31: 333–367.

Frölich, Markus 2007: Propensity score matching without conditional independence assumption—with an application to the gender wage gap in the United Kingdom. Econometrics Journal 10: 359–407.

Frosini, Benito V. 2006: Causality and causal models: A conceptual perspective. International Statistical Review 74: 305–334.

Gad, Ahmed M., Abeer S. Ahmed 2006: Analysis of longitudinal data with intermittent missing values using the stochastic EM algorithm. Computational Statistics & Data Analysis 50: 2702–2714.

Gad, Ahmed M., Abeer S. Ahmed 2007: Sensitivity analysis of longitudinal data with intermittent missing values. Statistical Methodology 4: 217–226.

Gaetan, Carlo, Jian-Feng Yao 2003: A multiple-imputation Metropolis version of the EM algorithm. Biometrika 90: 643–654.

Gakidou, Emmanuela, Gary King 2006: Death by survey: Estimating adult mortality without selection bias from sibling survival data. Demography 43: 569–589.

Galavotti, Maria Carla, Patrick Suppes, Domenico Constantini (eds) 2001: Stochastic Causality. Stanford: CSLI Lecture Notes 131.

Galbraith, Jane I. 1991: The interpretation of a regression coefficient. Biometrics 47: 1593–1596.

Galler, Heinz-Peter, Ulrich Pötter 1992: Zur Robustheit von Schätzmodellen für Ereignisdaten. Pp. 379–405 in: Hujer, R., H. Schneider, W. Zapf (eds) 1992: Herausforderungen an den Wohlfahrtsstaat im strukturellen Wandel. Frankfurt: Campus.

Galles, D., Judea Pearl 1998: An axiomatic characterization of causal counterfactuals. Foundations of Science 3: 151–182.

Galton, A. 1984: The Logic of Aspect. Oxford: Clarendon Press.

Gallop, Robert J., Thomas R. ten Have, Paul Crits-Christoph 2006: A mixed effects Markov model for repeated binary outcomes with non-ignorable dropout. Statistics in Medicine 25: 2398–2426.

Gangl, Markus, Thomas A. DiPrete 2006: Kausalanalyse durch Matchingverfahren. Kölner Zeitschrift für Soziologie und Sozialpsychologie, Sonderheft 44 (Andreas Diekmann eds): 396–420.

Gao, Guozhi, Anastasios A. Tsiatis 2005: Semiparametric estimators for the regression coefficients in the linear transformation competing risks model with missing cause of failure. Biometrika 92: 875–891.

Garcia, Luis David, Michael Stillman, Bernd Sturmfels 2005: Algebraic geometry of Bayesian networks. Journal of Symbolic Computation 39: 331–355.

Gelman, Andrew, Xiao-Li Meng (eds) 2004: Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives. An Essential Journey with Donald Rubin's Statistical Family. New York: Wiley.

Gelman, Andrew, Iven van Mechelen, Geert Verbeke, Daniel F. Heitjan, Michel Meulders 2005: Multiple imputation for model checking: Complete-data plots with missing and latent data. Biometrics 61: 74–85.

Geneletti, Sara 2007: Identifying direct and indirect effects in a non-counterfactual framework. Journal of the Royal Statistical Society B 69: 199–215.

Gentleman, Robert, Alain C. Vandal 2002: Nonparametric estimation of the bivariate CDF for arbitrarily censored data. Canadian Journal of Statistics 30: 557–571.

Geraci, Marco, Matteo Bottai 2006: Use of auxiliary data in semi-parametric spatial regression with non-ignorable missing responses. Statistical Modelling 6: 321–336.

Ghosh, J.K., N.L. Hjort, C. Messan, R.V. Ramamoorthi 2006: Bayesian bivariate survival estimation. Journal of Statistical Planning and Inference 136: 2297–2308.

Gijbels, I., N. Veraverbeke 1991: Almost sure asymptotic representation for a class of functionals of the Kaplan–Meier estimator. Annals of Statistics 19: 1457–1470.

Gill, Richard D. 1983: Large sample behaviour of the product–limit estimator on the whole line. Annals of Statistics 11: 49–58.

Gill, Richard D. 1989: Non- and semiparametric MLE and the von Mises method I. Scandinavian Journal of Statistics 16: 97–128.

Gill, Richard D. 1992: Multivariate survival analysis. Theory of Probability and its Applications 37: 18–31, 284–301.

Gill, Richard D. 1994: Lectures on survival analysis. Pp. 115–241 in: D. Bakry, Richard D. Gill, S.A. Molchanov (eds) 1994: Lectures on Probability Theory. Ecole d'Eté de Probabilités de Saint–Flour XXII. Berlin: Springer.

Gill, Richard D. 1997: Nonparametric estimation under censoring and passive registration. Satistica Neerlandica 51: 35–54.

Gill, Richard D. 2005: Notes on CAR. http://www.math.uu.nl/people/gill/Onderwijs/Missing/carnote.pdf

Gill, Richard D., Peter D. Grünwald 2005: An algorithmic and a geometric characterization of Coarsening at Random. arXiv:math.ST/0510276 (last version: 2007-09-13).

Gill, Richard D., James M. Robins 1997: Sequential models for coarsening and missingness. Pp. 295–305 in: D.Y. Lin, Thomas R. Fleming (eds) 1997: Proceedings of the 1st Seattle Symposium in Biostatistics: Survival Analysis. Berlin: Springer.

Gill, Richard D., Yehuda Vardi, Jon A. Wellner 1988: Large sample theory of empirical distributions in biased sampling models. Annals of Statistics 16: 1069–1112.

Gill, Richard D., Mark van der Laan, John A. Wellner 1995: Inefficient estimators of the bivariate survival function. Annales de l'Institut Henri Poincaré, Probabilités et Statistiques 31: 545–597.

Gill, Richard D., Aad D. van der Vaart 1993: Non- and semi-parametric maximum likelihood estimators and the von Mises method II. Scandinavian Journal of Statistics 20: 271–288.

Gill, Richard D., Mark J. van der Laan, James M. Robins 1997: Coarsening at random: Characterizations, conjectures, counter-examples. Pp. 255–294 in: D.Y. Lin, Thomas R. Fleming (eds) 1997: Proceedings of the 1st Seattle Symposium in Biostatistics: Survival Analysis. Berlin: Springer.

Gleser, Leon Jay, Jiunn T. Hwang 1987: The nonexistence of 100(1-α)% confidence sets of finite diameter in errors–in–variables and related models. Annals of Statistics 15, 1351–1362.

Glickman, Mark E., Sharon-Lise T. Normand 2000: The derivation of a latent threshold instrumental variables model. Statistica Sinica 10: 517–544.

Goffinet, B. 1987: Alternative conditions for ignoring the process that causes missing data. Biometrika 74: 437–439.

Golan, Amos, Enrico Moretti, Jeffrey M. Perloff 2004: A small sample estimator for the sample-selection model. Econometric Reviews 23: 71–91.

Goldenberg, S. 1998: Rediscovering and confronting critical ambiguities in the determination of causality. Quality & Quantity 32: 181–200.

Goldenshluger, A.V., B.T. Polyak, B.T. 1993: Estimation of regression parameters with arbitrary noise. Mathematical Methods in Statistics 2: 18–29.

González-Manteiga, Wenceslao, Ana Pérez-González 2006: Goodness-of-fit tests for linear regression models with missing response data. Canadian Journal of Statistics 34: 149–170.

Goodman, Nelson 1976^2: Languages of Art. An Approach to a Theory of Symbols. Indianapolis: Hackett.

Gørgens, Tue 2003: Semi-parametric estimation of censored transformation models. Nonparametric Statistics 15: 377–393.

Gould, A., J.F. Lawless 1988: Consistency and efficiency of regression coefficient estimates in location-scale models. Biometrika 75: 535–540.

Gourieroux, C., A. Montfort, A. Trognon 1984: Pseudo maximum likelihood methods: Theory. Econometrica 52: 681–700.

Grayling, A.C. 1982: An Introduction to Philosophical Logic. Brighton: Harvester Press.

Granger, C.W.J. 1969: Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37: 424–438.

Greenland, Sander 2004: An overview of methods for causal inference from observational studies. Pp. 3–13 in: Gelman, Andrew, Xiao-Li Meng (eds) 2004: Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives. An Essential Journey with Donald Rubin's Statistical Family. New York: Wiley.

Greenland, Sander 2005a: Multiple-bias modelling for analysis of observational data (with discussion). Journal of the Royal Statistical Society A 168: 267–306.

Greenland, Sander 2005b: Epidemiologic measures and policy formulation: Lessons from potential outcomes. Emerging Themes in Epidemiology 2: 5. Discussion: Vol. 2: 4, 2: 3, 2: 2.

Greenland, Sander, William D. Finkle 1995: A critical look at methods for handling missing covariates in epidemiologic regression analyses. American Journal of Epidemiology 142: 1255–1264.

Greenland, Sander, James M. Robins, Judea Pearl 1999: Confounding and collapsibility in causal inference. Statistical Science 14: 29-46.

Greenland, Sander, Babette Brumback 2002: An overview of relations among causal modelling methods. International Journal of Epidemiology 31: 1030–1037.

Greenwood, P.E., Wolfgang Wefelmeyer 1998: Cox's factoring of regression model likelihoods for continuous-time processes. Bernoulli 4: 65–80.

Groeneboom, Piet, Jon A. Wellner 1992: Information Bounds and Nonparametric Maximum Likelihood Estimation. Basel: Birkhäuser.

Grünwald, Peter D., Joseph Y. Halpern 2003: Updating probabilities. Journal of Artificial Intelligence Research 19: 243–278.

Gustafson, Paul 2001: On measuring sensitivity to parametric model misspecification. Journal of the Royal Statistical Society B 63: 81–94.

Haavelmo, Trygve 1944: The Probability Approach to Econometrics. Econometrica 12, Supplement: 1–115.

Hacker, P.M.S. 1982: Events and objects in space and time. Mind 91: 1–19.

Hacking, Ian 1988: Telepathy: Origins of randomization in experimental design. Isis 79: 427–451.

Hájek, Alan 2003: What conditional probability could not be. Synthese 137: 273–323.

Hall, Ned 2004: Rescued from the rubbish bin: Lewis on causation. Philosophy of Science 71: 1107–1114.

Hall, Ned 2007: Structural equations and causation. Philosophical Studies 132: 109–136.

Halpern, Joseph Y., Judea Pearl 2005: Causes and explanations: A structural-model approach. Part I: Causes. Part II: Explanations. British Journal for the Philosophy of Science 56: 843–887, 889–911.

Han, A.K. 1987: Non-parametric analysis of a generalized regression model. Journal of Econometrics 35: 303–316.

Harding, David J. 2003: Counterfactual models of neighborhood effects: The effect of neighborhood poverty on dropping out and teenage pregnancy. American Journal of Sociology 109: 676–719.

Harel, Ofer, Xia-Hua Zhou 2007: Multiple imputation: Review of theory, implementation and software. Statistics in Medicine 26: 3057–3077.

Hart, Andrew, Heinrich Matzinger 2006: Markers for error-corrupted observations. Stochastic Processes and their Applications 116: 807–829.

Hartfiel, Darald J. 1998: Markov Set-Chains. Berlin: Springer.

Hansen, Morris H., William G. Madow, Benjamin J. Tepping 1983: An evaluation of model-dependent and probability-sampling inferences in sample surveys (eith discussion). Journal of the American Statistical Association 78: 776–807.

Hattori, Satoshi 2006: Some properties of misspecified additive hazards models. Statistics & Probability Letters 76: 1641–1646.

Hausman, Daniel M., James Woodward 1999: Independence, invariance and the causal Markov condition. British Journal for the Philosophy of Science 50: 521–583.

Hausman, Daniel M., James Woodward 2004: Modularity and the causal Markov condition: A Restatement. British Journal for the Philosophy of Science 55: 147–161.

Hausman, Daniel M. 2005: Causal relata: Tokens, Types, or Variables? Erkenntnis 63: 33–54.

Hawkes, Denise, Ian Plewis 2006: Modelling non-response in the National Child Development Study. Journal of the Royal Statistical Society A 169: 479–491.

Haziza, David, J.N.K. Rao 2005: Inference for domains under imputation for missing survey data. Canadian Journal of Statistics 33: 149–161.

He, Wenqin, Jerald F. Lawless 2005: Bivariate location-scale models for regression analysis, with applications to life time data. Journal of the Royal Statistical Society B 67: 63–78.

Heckman, James J. 1976: The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models. Annals of Economic and Social Measurement 5: 475–492.

Heckman, James J. 1979: Sample selection bias as a specification error. Econometrica 47: 153–161.

Heckman, James J. 1990: Varieties of selection bias. American Economic Review 80: 313–318.

Heckman, James J. 1997: Instrumental variables. A study of implicit behavioral assumptions used in making program evaluations. The Journal of Human Resources 32: 441–462.

Heckman, James J. 2000: Causal parameters and policy analysis in economics: A twentieth century retrospective. The Quarterly Journal of Economics 115: 45–97.

Heckman, James J. 2001: Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture. Journal of Political Economy 109: 673–748.

Heckman, James J. 2006: The scientific model of causality (with discussion). Sociological Methodology 35: 1–162.

Heckman, James J. 2008: Econometric causality. International Statistical Review 76: 1–27.

Heckman, James J., Bo E. Honoré 1990: The empirical content of the Roy model. Econometrica 58: 1121–1149.

Heckman, James J., V. Joseph Hotz 1989: Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training. Journal of the American Statistical Association 84: 862–880.

Heckman, James J., Hidehiko Ichimura, Jeffrey Smith, Petra Todd 1996: Sources of selection bias in evaluating social programs: An interpretation of conventional measures and evidence on the effectiveness of matching as a program evaluation method. Proceedings of the National Academy of Sciences 93: 13416–13420.

Heckman, James J., Hidehiko Ichimura, Petra Todd 1997: Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. Review of Economic Studies 64: 605–654.

Heckman, James J., Salvador Navarro 2007: Dynamic discrete choice and dynamic treatment effects. Journal of Econometrics 136: 341–396.

Heckman, James J., Jeffrey Smith (with the assistance of Nancy Clements) 1997: Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. Review of Economic Studies 64: 487–535.

Heckman, James J., Sergio Urzua, Edward Vytlacil 2006: Understanding instrumental variables in models with essential heterogeneity. Review of Economics and Statistics 88: 389–432.

Heckman, James J., Edward Vytlacil 1998: Instrumental variables methods for the correlated random coefficient model. The Journal of Human Resources 33: 974–987.

Heckman, James J., Edward Vytlacil 1999: Local instrumental variables and latent variable models for identifying and bounding treatment effects. Proceedings of the National Academy of Sciences 96: 4730–4734.

Heckman, James J., Edward Vytlacil 2005: Structural equations, treatment effects, and econometric policy evaluation. Econometrica 73: 669–738.

Hedström, Peter, Richard Swedberg (eds) 1998: Social Mechanisms. An Analytical Approach to Social Theory. Cambridge: Cambridge University Press.

Heitjan, Daniel F. 1993: Ignorability and coarse data: Some biomedical examples. Biometrics 49: 1099–1109.

Heitjan, Daniel F. 1994: Ignorability in general incomplete-data models. Biometrika 81: 701–708.

Heitjan, Daniel F., Donald B. Rubin 1991: Ignorability and coarse data. The Annals of Statistics 19: 2244–2253.

Heitjan, Daniel F., Srabashi Basu 1996: Distinguishing "Missing at Random" and "Missing Completely at Random". American Statistician 50: 207–213.

Heller, Glenn 2007: Smoothed rank regression with censored data. Journal of the American Statistical Association 102: 552–559.

Helmers, Roelof, I. Wayan Mangku, Ričardas Zitikis 2007: A non-parametric estimator for the doubly periodic Poisson intensity function. Statistical Methodology 4: 481–492.

Hendry, David F., Edward E. Leamer, Dale J. Poirier 1990: A conversation on econometric methodology. Econometric Theory 6: 171-261.

Henmi, Masayuki 2004: A paradoxical effect of nuisance parameters on efficiency of estimators. Journal of the Japan Statistical Society 34: 75–86.

Henmi, Masayuki, Shinto Eguchi 2004: A paradox concerning nuisance parameters and projected estimating functions. Biometrika 91: 929–941.

Hennig, Christian 2007: Falsification of propensity models by statistical tests and the goodness-of-fit paradox. Philosophia Mathematica 15: 166–192.

Henriques, Carla, Paulo Eduardo Oliveira 20003: Estimation of a two-dimensional distribution function under association. Journal of Statistical Planning and Inference 113: 137–150.

Hens, Niel, Marc Aerts, Geert Molenberghs, Herbert Thijs, Geert Verbeke 2004: Kernel weighted influence measures. Computational Statistics & Data Analysis 48: 467–487.

Hens, Niel, Marc Aerts, Geert Molenberghs 2006: Model selection for incomplete and design-based samples. Statistics in Medicine 25: 2502–2520.

Hernán, Miguel A. 2004: A definition of causal effect for epidemiological research. Journal of Epidemiology and Community Health 58: 265–271.

Hernán, Miguel A., Sonia Hernández-Díaz, James M. Robins 2004: A structural approach to selection bias. Epidemiology 15: 615–625.

Hernán, Miguel A., Stephen R. Cole, Joseph Margolick, Mardge Cohen, James M. Robins 2005: Structural accelerated failure time models for survival analysis with time varying treatments. Pharmacoepidemiology and Drug Safety 14: 477–491.

Hernán, Miguel A., Emilie Lanoy, Dominique Costagliola, James M. Robins 2006: Comparison of dynamic treatment regimes via inverse probability weighting. Basic & Clinical Pharmacology & Toxicology 98: 237–242.

Hernán, Miguel A., James M. Robins 2006a: Instruments for causal inference. An Epidemiologist's dream? Epidemiology 17: 360–372.

Hernán, Miguel A., James M. Robins 2006b: Estimating causal effects from epidemiological data. Journal of Epidemiology and Community Health 60: 578–586.

Herring, Amy H., Joseph G. Ibrahim, Stuart R. Lipsitz 2004: Non-ignorable missing covariate data in survival analysis: A case-study of an international breast cancer study group trial. Applied Statistics 53: 293–310.

Heuchenne, Cédric, Ingrid van Keilegom 2007a: Polynomial regression with censored data based on preliminary nonparametric estimation. Annals of the Institute of Statistical Mathematics 59: 273–297.

Heuchenne, Cédric, Ingrid van Keilegom 2007b: Location estimation in nonparametric regression with censored data. Journal of Multivariate Analysis 98: 1558–1582.

Heyde, Christopher C. 1997: Quasi-Likelihood and its Application. A General Approach to Optimal Parameter Estimation. Berlin: Springer.

Hiddleston, Eric 2005a: A causal theory of counterfactuals. Noûs 39: 632–657.

Hiddleston, Eric 2005b: Causal powers. British Journal of the Philosophy of Science 56: 27–59.

Hines, R.J. O'Hara, W.G.S. Hines 2006: Letter to the editor: "An index of local sensitivity to nonignorable drop-out in longitudinal modelling" by Ma et al. Statistics in Medicine 25: 3217–3223.

Hines, R.J. O'Hara, W.G.S. Hines 2007: Covariance miss-specification and the local influence approach in sensitivity analyses of longitudinal data with drop-outs. Computational Statistics & Data Analysis 51: 5537–5546.

Hirano, Keisuke, Guido W. Imbens 2004: The propensity score with continuous treatments. Pp. 73–84 in: Gelman, Andrew, Xiao-Li Meng (eds) 2004: Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives. An Essential Journey with Donald Rubin's Statistical Family. New York: Wiley.

Hirano, Keisuke, Guido W. Imbens, Geert Ridder 2003: Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71: 1161–1189.

Hirano, Keisuke, Guido W. Imbens, Geert Ridder, Donald B. Rubin 2001: Combining panel data sets with attrition and refreshment samples. Econometrica 69: 1645–1659.

Hitomi, Kohtaro, Yoshihiko Nishiyama 2005: A paradox of semiparametric estimators with infinite dimensional nuisance parameters. Pp. 821–827 in: Zerger, A., R.M. Argent (eds) 2005: MODSIM 2005 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand, December 2005. http://www.mssanz.org.au/modsim05/papers/hitomi.pdf.

Hjort, Niels L. 1992: On inference in parametric survival data models. International Statistical Review 60: 355–387.

Hoffmann-Jørgensen, J. 1994: Probability with a View toward Statistics. Vol. I, Vol. II. London: Chapman & Hall.

Höfler, Michael 2005a: Causal inference based on counterfactuals. BMC Medical Research Methodology 5: 28.

Höfler, Michael 2005b: The Bradford Hill considerations on causality: A counterfactual perspective. Emerging Themes in Epidemiology 2: 11.

Höfler, Michael 2006: Getting causal considerations back on the right track. Emerging Themes in Epidemiology 3: 8.

Hogan, Joseph W., Joo Yeon Lee 2004: Marginal structural quantile models for longitudinal observational studies with time-varying treatment. Statistica Sinica 14: 927–944.

Hogan, Joseph W., Xihong Lin, Benjamin Herman 2004: Mixtures of varying coefficient models for longitudinal data with discrete or continuous nonignorable dropout. Biometrics 60: 854–864.

Hogan, Joseph W., Tony Lancaster 2004: Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. Statistical Methods in Medical Research 13: 17–48.

Hogan, Joseph W., Jason Roy, Christina Korkontzelou 2004: Handling drop-out in longitudinal studies. Statistics in Medicine 23: 1455–1497.

Holbrook, John A.R. 1981: Stochastic independence and space-filling curves. American Mathematical Monthly 88: 426–432.

Holland, Paul W. 1986: Statistics and causal inference (with discussion). Journal of the American Statistical Association 81: 945–970.

Honoré, Bo E., Elie Tamer 2006: Bounds on parameters in panel dynamic discrete choice models. Econometrica 74: 611–629.

Hoover, Kevin D. 2001: Causality in Macroeconomics. Cambridge: Cambridge University Press.

Hornsteiner, Ulrich, Alfred Hamerle 1996: A combined GEE/Buckley-James method for estimating an accelerated failure time model of multivariate failure times. Discussion Paper 47, Sfb386, München.

Hornsteiner, Ulrich, Alfred Hamerle, Paul Michels, 1997: Parametric vs. nonparametric treatment of unobserved heterogeneity in multivariate failure times. Discussion Paper 80, Sfb386, München.

Hornsteiner, Ulrich 1998: Statistische Analyse multivariater Ereignisdaten mit Anwendungen in der Werbewirkungsforschung und in der Kardiologie. Dissertation, Regensburg.

Horowitz, Joel L., Charles F. Manski 1998: Censoring of outcomes and regressors due to survey nonresponse: Identification and estimation using weights and imputations. Journal of Econometrics 84: 37–58.

Horowitz, Joel L., Charles F. Manski 2000: Nonparametric analysis of randomized experiments with missing covariate and outcome data (with discussion). Journal of the American Statistical Association 95: 77–88.

Horowitz, Joel L., Charles F. Manski 2006: Identification and estimation of statistical functionals using incomplete data. Journal of Econometrics 132: 445–459.

Horowitz, Joel L., Charles F. Manski, Maria Ponomareva, Jörg Stoye 2003: Computation of bounds on population parameters when the data are incomplete. Reliable Computing 9: 419–440.

Horton, Nicholas J., Garrett M. Fitzmaurice 2002: Maximum likelihood estimation of bivariate logistic models for incomplete responses with indicators of ignorable and non-ignorable missingness. Applied Statistics 51: 281–295.

Horton, Nicholas J., Ken P. Kleinman 2007: Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. American Statistician 61: 79–90.

Horton, Nicholas J., Stuart R. Lipsitz 2001: Multiple imputation in practice: Comparison of software packages for regression models with missing variables. The American Statistician 55: 244–254.

Horton, Nicholas J., Stuart R. Lipsitz, Michael Parzen 2003: A potential bias when rounding in multiple imputation. The American Statistician 57: 229–232.

Horton, Nicholas J., Nan M. Laird 1998: Maximum likelihood analysis of generalized linear models with missing covariates. Statistical Methods in Medical Research 8: 37–50.

Howson, Colin 2000: Hume's Problem: Induction and the Justification of Belief. Oxford: Oxford University Press.

Hsu, Chiu-Hsieh, Jeremy M.G. Taylor, Susan Murray, Daniel Commenges 2006: Survival analysis using auxiliary variables via non-parametric multiple imputation. Statistics in Medicine 25: 3503–3517.

Hsu, Chiu-Hsieh, Jeremy M.G. Taylor, Susan Murray, Daniel Commenges 2007: Multiple imputation for interval censored data with auxiliary variables. Statistics in Medicine 26: 769–781.

Hu, X. Joan, Stephen W. Lagakos 2007: Nonparametric estimation of the mean function of a stochastic process with missing observations. Lifetime Data Analysis 13: 51–73.

Hu, X. Joan, R. Jason Schroeder, Winfred C. Wang, James M. Boyett 2007: Pseudoscore-based estimation from biased observations. Statistics in Medicine 26: 2836–2852.

Huang, Jian, Shuangge Ma, Huiliang Xie 2007: Least absolute deviations estimation for the accelerated failure time model. Statistica Sinica 17: 1533–1548.

Huang, Jie, David Harrington 2004: Dimension reduction in the linear model for right-censored data: Predicting the change of HIV-I RNA levels using clinical and protease gene mutation data. Lifetime Data Analysis 10: 425–443.

Huang, Jie, David Harrington 2005: Iterative partial least squares with right-censored data analysis: A comparison to other dimension reduction techniques. Biometrics 61: 17–24.

Huang, Lan, Ming-Hui Chen, Joseph G. Ibrahim 2005: Bayesian analysis for generalized linear models with nonignorably missing covariates. Biometrics 61: 767–780.

Hubbard, Alan E., Mark J. van der Laan 2008: Population intervention models in causal inference. Biometrika 95: 35–47.

Huber, Peter J. 1967: The behavior of maximum likelihood estimates under nonstandard conditions. Pp. 221–233 in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Berkley: University of California Press.

Huber, Peter J. 1976: Kapazitäten statt Wahrscheinlichkeiten? Gedanken zur Grundlegung der Statistik. Jahresberichte der Deutschen Mathematiker-Vereinigung 78: 81–92.

Huber, Peter J. 1981: Robust Statistics. New York: Wiley.

Hudgens, Michael G. 2005: On nonparametric maximum likelihood estimation with interval censoring and left truncation. Journal of the Royal Statistical Society B 67: 573–587.

Humphreys, Paul 2008: Probability theory and its models. Pp. 1–11 in: Nolan, Deborah, Terry Speed (eds) 2008: Probability and Statistics: Essays in Honor of David A. Freedman, vol 2. Beachwood, Ohio: Institute of Mathematical Statistics.

Hutter, Marcus, Marco Zaffalon 2004: Distribution of mutual information from complete and incomplete data. Computational Statistics & Data Analysis 48: 633–657.

Ibragimov, I.A., R.Z. Has’minskii 1981: Statistical Estimation: Asymptotic Theory. Berlin: Springer.

Ibrahim, Joseph G., Ming-Hui Chen, Stuart R. Lipsitz 2001: Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. Biometrika 88: 551–564.

Ichimura, Hidehiko, Christopher Taber 2001: Propensity-score matching with instrumental variables. American Economic Review 91: 119–124.

Imai, Kosuke, David A. van Dyk 2004: Causal inference with general treatment regimes: Generalizing the propensity score. Journal of the American Statistical Association 99: 854–866.

Imai, Kosuke 2008: Sharp bounds on the causal effects in randomized experiments with "truncation-by-death". Statistics & Probability Letters 78: 144–149.

Imbens, Guido W., Charles F. Manski 2004: Confidence intervals for partially identified parameters. Econometrica 72: 1845–1857.

Imbens, Guido W., Paul W. Rosenbaum 2005: Robust, accurate confidence intervals with a weak instrument: quarter of birth and education. Journal of the Royal Statistical Society A 168: 109–126.

Imbens, Guido W., Thomas Lemieux 2008: Regression discontinuity designs: A guide to practice. Journal of Econometrics 142: 615–635.

Ingram, D.D., J.C. Kleinman 1989: Empirical comparisons of proportional hazards and logistic regression models. Statististics in Medicine 8: 525–538.

Ivanoff, B. Gail, Ely Merzbach 2004: Random clouds and an application to censoring and survival analysis. Stochastic Processes and their Applications 111: 259–279.

Jackson, Dan, John Copas, Alex J. Sutton 2005: Modelling reporting bias: The operative mortality rate for ruptured abdominal aortic aneurysm repair. Journal of the Royal Statistical Society A 168: 737–752.

Jacobsen, Martin, Niels Keiding 1995: Coarsening at random in general sample spaces and random censoring in continuous time. The Annals of Statistics 23: 774–786.

Jaeger, Manfred 2005a: Ignorability for categorical data. Annals of Statistics 33: 1964–1981.

Jaeger, Manfred 2005b: Ignorability in statistical and probabilistic inference. Journal of Artificial Intelligence Research 24: 889–917.

Jaeger, Manfred 2005c: A logic for inductive probabilistic reasoning. Synthese 144: 181–248.

Jaeger, Manfred 2006: On testing the missing at random assumption. Pp. 671–678 in: Johannes Fürnkranz, Tobias Scheffer, Myra Spiliopoulou (eds) 2006: Machine Learning: ECML 2006. Proceedings of the 17th European Conference on Machine Learning, Berlin, Germany, September 18-22, 2006. Lecture Notes on Computer Science 4212. Berlin: Springer.

Jagers, Peter 1989: The Markov structure of population growth. Acta Applicandae Mathematicae 14: 103–114.

Jansen, Ivy, Niel Hens, Geert Molenberghs, Marc Aerts, Geert Verbeke, Michael G. Kenward 2006: The nature of sensitivity in monotone missing not at random models. Computational Statistics & Data Analysis 50: 830–858.

Jansen, Ivy, Caroline Beunckens, Geert Molenberghs, Geert Verbeke, Craig Mallinckrodt 2006: Analyzing incomplete discrete longitudinal clinical trial data. Statistical Science 21: 52–69.

Jaynes, Edwin T. 2003: Probability Theory. The Logic of Science. Cambridge: Cambridge University Press.

Jeffrey, Richard 2004: Subjective Probability: The Real Thing. Cambridge: Cambridge University Press.

Jemiai, Yannis, Andrea Rotnitzky, Bryan E. Shepherd, Peter B. Gilbert 2007: Semiparametric estimation of treatment effects given base-line covariates on an outcome measured after a post-randomization event occurs. Journal of the Royal Statistical Society B 69: 879–901.

Jiang, Hongyu, Jason P. Fine, Michael R. Kosorok, Rick Chappell 2005: Pseudo self-consistent estimation of a coopula model with informative censoring. Scandinavian Journal of Statistics 32: 1–20.

Jin, Zhezhen 2007: M-estimation in regression models for censored data. Journal of Statistical Planning and Inference 137: 3894–3903.

Jin, Zhezhen, D.Y. Lin, L.J. Wei, Zhiliang Ying 2003: Rank-based inference for the accelerated failure time model. Biometrika 90: 341–353.

Jin, Zhezhen, D.Y. Lin, Zhiliang Ying 2006: On least-squares regression with censored data. Biometrika 93: 147–161.

Jin, Zhezhen, D.Y. Lin, Zhiliang Ying 2006a: Rank regression analysis of multivariate failure time data based on marginal linear models. Scandinavian Journal of Statistics 33: 1–23.

Jin, Zhezhen, Zhiliang Ying, L.J. Wei 2001: A simple resampling method by perturbing the minimand. Biometrika 88: 381–390.

Joffe, Marshall M., Paul R. Rosenbaum 1999: Propensity scores. American Journal of Epidemiology 150: 327–333.

Johnson, Brent A. 2008: Variable selection in semiparametric linear regression with censored data. Journal of the Royal Statistical Society B 70: 351–370.

Johnson, Brent A., Dennis D. Boos 2005: A note on the use of kernel functions in weighted estimators. Statitics & Probability Letters 72: 345–355.

Johnson, Brent A., Anastasios A. Tsiatis 2005: Semiparametric inference in observational duration-response studies, with duration possibly right-censored. Biometrika 92: 605–618.

Jones, Andrew M., Xander Joolman, Nigel Rice 2006: Health-related non-response in the British Household Panel Survey and European Community Household Panel: Using inverse-probability-weighted estimators in non-linear models. Journal of the Royal Statistical Society A 169: 543–569.

Jonker, Marianne A. 2003: Estimation of life expectancy in the Middle Ages. Journal of the Royal Statistical Society A 166: 105–117.

Jonker, Marianne A., Aad van der Vaart 2005: Estimation of average mortality under censoring and truncation. Journal of Population Research 22: 49–62.

Judkins, David R., David Morganstein, Paul Zador, Andrea Piesse, Brandon Barrett, Pushpal Mukhopadhyay 2007: Variable selection and raking in propensity scoring. Statistics in Medicine 26: 1022–1033.

Kac, Mark 1959: Statistical Independence in Probability, Analysis and Number Theory. New York: Wiley.

Kahn, Shakeeb, Elie Tamer 2007: Partial rank estimation of duration models with general forms of censoring. Journal of Econometrics 136: 251–280.

Kalbfleisch, John D., R.J. MacKay 1979: On constant sum models for censored survival data. Biometrika 66: 87–90.

Kalbfleisch, John D., Ross L. Prentice 2002^2: Statistical Analysis of Failure Time Data. New York: Wiley.

Kalton, Graham 2002: Models in the practice of survey sampling (revisited) (with discussion). Journal of Official Statistics 18: 129–161. ls

Kalton, Graham, Ismael Flores-Cervantes 2003: Weighting methods. Journal of Official Statistics 19: 81–97.

Kalton, Graham, Andrea Piesse 2007: Survey research methods in evaluation and case-control studies. Statistics in Medicine 26: 1675–1687.

Kang, Joseph S.Y., Joseph L. Schafer 2007: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). Statistical Science 22: 523–580.

Karr, Alan F. 1991^2: Point Processes and Their Statistical Inference. New York: Marcel Dekker.

Kass, Robert E., Paul W. Vos 1997: Geometrical Foundations of Asymptotic Inference. New York: Wiley.

Kateri, M., N. Balakrishnan 2008: Statistical evidence in contingency tables analysis. Journal of Statistical Planning and Inference 138: 873–887.

Keiding, Niels 1999: Event history analysis and inference from observational epidemiology. Statistics in Medicine 18: 2353–2363.

Keiding, Niels 2006: Event history analysis and the cross-section. Statistics in Medicine 25: 2343–2364.

Keiding, Niels, Richard D. Gill 1990: Random truncation models and Markov processes. Annals of Statistics 18: 582–602.

Keiding, Niels, Per Kragh Andersen, John P. Klein 1997: The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. Statistics in Medicine 16: 215–224.

Kelsen, Hans 1982: Vergeltung und Kausalität. Graz: Böhlau.

Kenna, Leslie A., Lewis B. Sheiner 2004: Estimating treatment effect in the presence of non-compliance measured with error: Precision and robustness of data analysis methods. Statistics in Medicine 23: 3561–3580.

Kenward, Michael G. 1998: Selection models for repeated measurements with non-random dropout: An illustration of sensitivity. Statistics in Medicine 17: 2723–2732.

Kenward, Michael G., James Carpenter 2007: Multiple imputation: Current perspectives. Statistical Methods in Medical Research 16: 199–218.

Kenward, Michael G., Gert Molenberghs, H. Thijs 2003: Pattern-mixture models with proper time dependence. Biometrika 90: 53–71.

Kim, J. 1969: Events and their descriptions: Some considerations. Pp. 198–215 in: Rescher, N. (ed) 1969: Essays in Honor of Carl G. Hempel. Dordrecht: D. Reidel.

Kim, Jae Kwang 2004: Finite sample properties of multiple imputation estimators. Annals of Statistics 32: 766–783.

Kim, Jae Kwang, Wayne Fuller 2004: Fractional hot deck imputation. Biometrika 91: 559–578.

Kim, Jae Kwang, J. Michael Brick, Wayne Fuller, Graham Kalton 2006: On the bias of the multiple-imputation variance estimator in survey sampling. Journal of the Royal Statistical Society B 68: 509–521.

Kim, Jae Kwang, Hyeonah Park 2006: Imputation using response probability. Canadian Journal of Statistics 34: 171–182.

Kim, Kyoo il 2006: Sample selection models with a common dummy endogeneous regressor in simultaneous equations: A simple two-step estimation. Economic Letters 91: 280–286.

Kim, Yang-Jin, Myoungshic Jhun 2008: Analysis of recurrent event data with incomplete observation gaps: Statistics in Medicine 27: 1075–1085.

King, Alan J. 1989: Generalized delta theorems for multivalued mapings and measurable selections. Mathematics of Operations Research 14: 720–736.

King, Gary, Langche Zeng 2006: The dangers of extreme counterfactuals. Political Analysis 14: 131–159.

Kingman, J.F.C. 1993: Poisson Processes. Oxford: Oxford University Press.

Klaassen, Chris A.J., Andries J. Lenstra 2003: Vanishing Fisher information. Acta Applicandae Mathematicae 78: 193–200.

Klain, Daniel A., Gian-Carlo Rota 1997: Introduction to Geometric Probability. Cambridge: Cambridge University Press.

Klein, Thomas 1993: Soziale Determinanten der Lebenserwartung. Kölner Zeitschrift für Soziologie und Sozialpsychologie 45: 712–730.

Kluve, Jochen 2004: On the role of counterfactuals in inferring causal effects. Foundations of Science 9: 65–101.

Knuuttila, Tarja, Atro Voutilainen 2003: A parser as an epistemic artefact: A material view on models. Philosophy of Science 70: 1484–1495.

Koch, Achim 1997: Teilnahmeverhalten beim ALLBUS 1994: Soziodemographische Determinanten von Erreichbarkeit, Befragungsfähigkeit und Kooperationsbereitschaft. Kölner Zeitschrift für Soziologie und Sozialpsychologie 49: 98–122.

Koch, Achim 2002: 20 Jahre Feldarbeit im ALLBUS: Ein Blick in die Blackbox. ZUMA Nachrichten 51: 9–37.

Kölling, Arnd, Susanne Rässler 2003: Die Einflüsse von Antwortverweigerung und mehrfacher Ergänzung fehlender Daten auf Produktivitätsschätzungen mit dem IAB-Betriebspanel. Jahrbücher für Nationalökonomie und Statistik 223: 279–311.

Kolmogorov, Andreĭ N. 1950: Foundations of the Theory of Probability. New York: Chelsea.

Kolmogorov, Andreĭ N. 1950: The theory of probability. Pp. 229–264 in: A.D. Aleksandrov, Andreĭ N. Kolmogorov, M.A. Lavrent'ev (eds) 1999: Mathematics. Its Content, Methods, and Meaning. Mineola: Dover.

Kong, Fanhui, Qiqing Yu 2007: Asymptotic distributions of the Buckley-James estimator under nonstandard conditions. Statistica Sinica 17: 341–360.

Korinek, Anton, Johan A. Mistiaen, Martin Ravallion 2007: An econometric method of correcting for unit nonresponse bias in surveys. Journal of Econometrics 136: 213–235.

Kott, Philip S. 1995: A paradox of multiple imputation. Proceedings of the Survey Research Methods Section, American Statistical Association 380–383.

Koul, H., V. Susarla, J. van Ryzin 1981: Regression analysis with randomly right censored data. Annals of Statistics 9: 1276–1288.

Kreinovich, Vladik, Scott Ferson 2006: Computing best-possible bounds for the distribution of a sum of several variables is NP-hard. International Journal of Approximate Reasoning 41: 331–342.

Kroh, Martin 2006: Taking 'Don't Knows' as valid responses: A multiple complete random imputation of missing data. Quality & Quantity 40: 225–244.

Kurland, Brenda F., Patrick J. Heagerty 2004: Marginalized transition models for longitudinal binary data with ignorable and non-ignorable drop-out. Statistics in Medicine 23: 2673–2695.

Kurland, Brenda F., Patrick J. Heagerty 2005: Directly parametrized regression conditioning on being alive: analysis of longitudinal data truncated by deaths. Biostatistics 6: 241–258.

Kuroki, Manabu 2007: Graphical identifiability criteria for causal effects in studies with an unobserved treatment/response variable. Biometrika 94: 37–47.

Kurth, Tobias, Alexander M. Walker, Robert J. Glynn, K. Arnold Chan, J. Michael Gaziano, Klaus Berger, James M. Robins 2005: Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. American Journal of Epidemiology 163: 262–270.

Laaksonen, Seppo 2003: Alternative imputation techniques for complex metric variables. Journal of Applied Statistics 30: 1009–1020.

Lad, Frank 1996: Operational Subjective Statistical Methods. A Mathematical, Philosophical, and Historical Introduction. New York: Wiley.

Lagakos, Stephen W. 1988a: The loss in efficiency from misspecifying covariates in proportional hazards regression models. Biometrika 75: 156–160.

Lagakos, Stephen W. 1988b: Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable. Statistics in Medicine 7: 257–274.

Lagakos, Stephen W., D.A. Schoenfeld 1984: Properties of proportional-hazards score tests under misspecified regression models. Biometrics 40: 1037–1048.

Lai, T.L., Z. Ying 1991: Large sample theory of a modified Buckley-James estimator for regression analysis with censored data. Annals of Statistics 19: 1370–1402.

Lai, T.L., Z. Ying 1994: A missing information principle and M-estimators in regression analysis with censored and truncated data. Annals of Statistics 22: 1222–1255.

Langberg, Naftali, Frank Proschan, A.J. Quinci 1978: Converting dependent models into independent ones, preserving essential features. Annals of Probability 6: 174–181.

Last, Günter, Andreas Brandt 1995: Marked Point Processes on the Real Line. The Dynamic Approach. Berlin: Springer.

Lawless, Jerald F. 2004: A note on interval-censored lifetime data and Oller et al.'s constant sum condition. Canadian Journal of Statistics 32: 327–331.

Lawless, Jerald F., John D. Kalbfleisch, Chris J. Wild 1999: Semiparametric methods for response-selective and missing data problems in regression. Journal of the Royal Statistical Society B 61: 413–438.

Lawley, D.N. 1943: A note on Karl Pearson's selection formulae. Proceedings of the Royal Society of Edinburgh A 62: 28–30.

Lazrieva, N.L., T.A. Toronjadze, T.A. 1991: On stable M-estimators in the partial likelihood scheme. Pp. 567–596 in: Sazonov, V.V., T. Shervashidze (eds) 1991: New Trends in Probability and Statistics. Utrecht: VSP.

Leamer, Edward E. 1983: Let's take the con out of econometrics. American Economic Review 73: 31–43.

Leamer, Edward E. 1985: Sensitivity analysis would help. American Economic Review 75: 308–313.

LeCam, Lucien, Grace Lo Yang 1988: On the preservation of LAN under information loss. Annals of Statistics 16: 483–520.

LeCam, Lucien, Grace Lo Yang 1990: Asymptotics in Statistics. Some Basic Concepts. Berlin: Springer.

Lechner, Michael 1999: Nonparametric bounds on employment and income effects of continuous vocational training in East Germany. Econometrics Journal 2: 1–28.

Lechner, Michael 2002: Some practical issues in the evaluation of heterogeneous labour market programmes by matching methods. Journal of the Royal Statistical Society A 165: 59–82.

Lechner, Michael 2008: A note on endogeneous control variables in causal studies. Statistics & Probability Letters 78: 190–195.

Lee, Byung-Joo, Lawrence C. Marsh 2000: Sample selection bias correction for missing response observations. Oxford Bulletin of Economics ans Statistics 62: 305–322.

Lee, Myoung-Jae 2004: Selection correction and sensitivity analysis for ordered treatment effect on count response. Journal of Applied Econometrics 19: 323–337.

Lee, Sik-Yum, Bin Lu, Xin-Yuan Song 2006: Assessing local influence for nonlinear structural equation models with ignorable missing data. Computational Statistics & Data Analysis 50: 1356–1377.

Lee, Sik-Yum, Xin-Yuan Song 2007: A unified maximum likelihood approach for analyzing structural equation models with missing nonstandard data. Sociological Methods & Research 35: 352–381.

Lee, Sik-Yum, Nian-Sheng Tang 2006: Analysis of nonlinear structural equation models with nonignorable missing covariates and ordered categorical data. Statistica Sinica 16: 1117–1141.

Leeb, Hannes, Benedikt M. Pötscher 2006: Can one estimate the conditional distribution of post-model-selection estimators? Annals of Statistics 34: 2554–2591.

Lehmann, Erich Leo 1990: Model specification: the views of Fisher and Neyman, and later developments. Statistical Science 5: 160–168.

Lehmann, Erich Leo 1999: Elements of Large-Sample Theory. Berlin: Springer.

Lehmann, Erich Leo, George Casella 1998^2: Theory of Point Estimation. Berlin: Springer.

Lehmann, Erich Leo, Wei-Yin Loh 1990: Pointwise versus uniform robustness of some large sample tests and confidence intervals. Scandinavian Journal of Statistics 17: 177–187.

Leng, Chenlei, Shuangge Ma 2007: Accelerated failure time models with nonlinear covariates effects. Australian & New Zealand Journal of Statistics 49: 155–172.

Lenhard, Johannes 2006: Models and statistical inference: The controversy between Fisher and Neyman-Pearson. British Journal for the Philosophy of Science 57: 69–91.

Lenstra, Andries J. 2005: Cramér–Rao revisited. Bernoulli 11: 263–282.

Leon, Andrew C., Donald Hedeker 2007: Quintile stratification based on a misspecified propensity score in longitudinal treatment effectiveness analyses of ordinal doses. Computational Statistics & Data Analysis 51: 6114–6122.

Leon, Selene, Anastasios A. Tsiatis, Marie Davidian 2003: Semiparametric estimation of treatment effect in a pretest-posttest study. Biometrics 59: 1046–1055.

Leung, Denis H.Y., Jing Qin 2006: Analysing survey data with incomplete responses by using a method based on empirical likelihood. Applied Statistics 55: 379–396.

Leung, Siu Fai, Shiti Yu 2000: Collinearity and two-step estimation of sample selection models: Problems, origins, and remedies. Computational Economics 15: 173–199.

Leurgans, S. 1987: Linear models, random censoring and synthetic data. Biometrika 74: 301–309.

Levy, Douglas E., A. James O´Malley, Sharon-Lise T. Normand 2004: Covariate adjustment in clinical trials with non-ignorable missing data and noncompliance. Statistics in Medicine 23: 2319–2339.

Lewbel, Arthur 2007: Estimation of average treatment effects with misclassification. Econometrica 75: 537–551.

Lewbel, Arthur, Susanne M. Schennach 2007: A simple ordered data estimator for inverse density weighted expectations. Journal of Econometrics 136: 189–211.

Li, Jingjin, Mark D. Schluchter 2004: Conditional mixed models adjusting for non-ignorable drop-out with administrative censoring in longitudinal studies. Statistics in Medicine 23: 3489–3503.

Li, Jinhui, Xiaowei Yang, Yingnian Wu, Steven Shoptaw 2007: A random effects Markov transition model for Poisson-distributed repeated measures with non-ignorable missing values. Statistics in Medicine 26: 2519–2532.

Li, Ker-Chau, Naihua Duan 1989: Regression analysis under link violation. Annals of Statistics 17: 1009–1052.

Li, Mingliang, Dale J. Poirier, Justin L. Tobias 2004: Do dropouts suffer from dropping out? Estimation and prediction of outcome gains in generalized selection models. Journal of Applied Econometrics 19: 203–225.

Li, Q.H., Stephen W. Lagakos 1997: Use of the Wei-Lin-Weissfeld method for the analysis of a recurring and a terminating event. Statistics in Medicine 16: 925–940.

Li, Xiaoming, Devan V. Mehrota, John Barnard 2006: Analysis of incomplete longitudinal binary data using multiple imputation. Statistics in Medicine 25: 2107–2124.

Liang, Hua 2008: Generalized partially linear models with missing covariates. Journal of Multivariate Analysis 99: 880–895.

Liang, Hua, Suojin Wang, Raymond J. Carroll 2007: Partially linear models with missing response variables and error-prone covariates. Biometrika 94: 185–198.

Liang, Hua, Suojin Wang, James M. Robins, Raymond J. Carroll 2004: Estimation in partially linear models with missing covariates. Journal of the American Statistical Association 99: 357–367.

Liang, Kung-Yee, Jing Qin 2000: Regression analysis under non-standard situations: a pairwise pseudolikelihood approach. Journal of the Royal Statistical Society B 62: 773–786.

Lin, D.Y., L.J. Wei 1989: The robust inference for the Cox proportional hazards model. Journal of the American Statistical Association 84: 1074–1078.

Lin, Haiqun, Charles E. McCulloch, Robert A. Rosenheck 2004: Latent pattern mixture models for informative intermittent missing data in longitudinal studies. Biometrics 60: 295–305.

Lin, Haiqun, Daniel O. Scharfstein, Robert A. Rosenheck 2004: Analysis of longitudinal data with irregular, outcome-dependent follow-up. Journal of the Royal Statistical Society B 66: 791–813.

Lin, J.S., L.J. Wei 1992: Linear regression analysis for multivariate failure time observations. Journal of the American Statistical Association 87: 1091–1097.

Lin, Nan, Xuming He 2006: Robust and efficient estimation under data grouping. Biometrika 93: 99–112.

Lindley, Dennis V. 2002: Seeing and doing: The concept of causation (with discussion). International Statistical Review 70: 191–214.

Lindvall, Torgny 1992: Lectures on the Coupling Method. Mineola: Dover.

Lipsitz, Stuart R., Joseph G. Ibrahim 2000: Estimation with correlated censored survival data with missing covariates. Biostatistics 1: 315–327.

Lipsitz, Stuart R., Geert Molenberghs, Garrett M. Fitzmaurice, Joseph G. Ibrahim 2004: Protective estimator for linear regressions with nonignorably missing Gaussian outcomes. Statistical Modeling 4: 3–17.

Lipsitz, Stuart R., Michael Parzen, Sundar Natarajan, Joseph G. Ibrahim, Garrett M. Fitzmaurice 2004: Generalized linear models with a coarsened covariate. Applied Statistics 53: 279–292.

Lipsitz, Stuart R., Lue Ping Zhao, Geert Molenberghs 1998: A semiparametric method of multiple imputation. Journal of the Royal Statistical Society B 60: 127–144.

Lipton, Robert, Terje Ødegaard 2005: Causal thinking and causal language in epidemiology: It's in the details. Epidemiologic Perspectives & Innovations: 2: 8. Discussion: vol. 3: 7.

Little, Roderick J.A. 1992: Regression with missing X's: A review. Journal of the American Statistical Association 87: 1227–1237.

Little, Roderick J.A. 1993: Pattern-mixture models for multivariate incomplete data. Journal of the American Statistical Association 88: 125–134.

Little, Roderick J.A., Hyonggin An 2004: Robust likelihood-based analysis of multivariate data with missing values. Statistica Sinica 14: 949–968.

Little, Roderick J.A., Fang Liu, Trivellore E. Raghunathan 2004: Statistical disclosure techniques based on multiple imputation. Pp. 141–152 in: Gelman, Andrew, Xiao-Li Meng (eds) 2004: Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives. An Essential Journey with Donald Rubin's Statistical Family. New York: Wiley.

Little, Roderick J., Trivellore Raghunathan 1997: Should imputation of missing data condition on all observed variables? Proceedings of the Survey Research Methods Section, American Statistical Association 617–622.

Little, Roderick J.A., Donald B. Rubin 2002^2: Statistical Analysis with Missing Data. New York: Wiley.

Little, Roderick J., Sonya Vartivarian 2003: On weighting the rates in non-response weights. Statistics in Medicine 22: 1589–1599.

Liu, Wei, Lang Wu 2007: Simultaneous inference for semiparametric nonlinear mixed-effects models with covariate measurement errors and missing responses. Biometrics 63: 342–350.

Lok, Judith J. 2007: Structural nested models and standard software: A mathematical foundation through partial likelihood. Scandinavian Journal of Statistics 34: 186–206.

Lok, Judith, Richard Gill, Aad W. van der Vaart, James Robins 2004: Estimating the causal effect of a time-varying treatment on time-to-event using structural nested failure time models. Statistica Neerlandica 58: 271–295.

Lombard, L.B. 1986: Events. A Metaphysical Study. London: Routledge & Kegan Paul,

Longford, Nicholas T., P. Tyer, U.A.M. Nur, H. Seivewright 2006: Analysis of a long-term study of neurotic disorder, with insights into the process of non-response. Journal of the Royal Statistical Society A 169: 507–523.

Lu, Bo 2005: Propensity score matching with time-dependent covariates. Biometrics 61: 721–728.

Lu, Guobing, John B. Copas 2004: Missing at random, likelihood ignorability and model completeness. Annals of Statistics 32: 754–765.

Lu, Kaifeng, Anastasios A. Tsiatis 2005: Comparison between two partial likelihood approaches for the competing risks model with missing cause of failure. Lifetime Data Analysis 11: 29–40.

Lu, Wenbin 2005: Marginal regression of multivariate event times based on linear transformation models. Lifetime Data Analysis 11: 389–404.

Lu, Wenbin, Yu Liang 2008: Analysis of competing risks data with missing cause of failure under additive hazards model. Statistica Sinica 18: 219–234.

Lu, Xuewen, Tsung-Lin Cheng 2007: Randomly censored partially linear single-index models. Journal of Multivariate Analysis 98: 1895–1922.

Lunceford, Jared K., Marie Davidian 2004: Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. Statistics in Medicine 23: 2937–2960.

Lund, Jens 2000: Sampling bias in population studies—How to use the Lexis diagram. Scandinavian Journal of Statistics 27: 589–604.

Luo, Xiaohui, Dennis D. Boos, Roy N. Tamura 2004: Score tests for dose effect in the presence of non-responders. Statistics in Medicine 23: 3581–3591.

Lübbe, W. 1994: Structural causes: On causal chains in social sciences. Pp. 91–108 in: Faye, J., U. Scheffler, M. Urchs (eds) 1994: Logic and Causal Reasoning. Berlin: Akademie Verlag.

Ma, Guoguang, Andrea B. Troxel, Daniel F. Heitjan 2005: An index of local sensitivity to nonignorable drop-out in longitudinal modelling. Statistics in Medicine 24: 2129–2150.

Ma, Guoguang, Daniel F. Heitjan 2004: Sensitivity to nonignorability in frequentist inference. Pp. 175–186 in: Gelman, Andrew, Xiao-Li Meng (eds) 2004: Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives. An Essential Journey with Donald Rubin's Statistical Family. New York: Wiley.

Ma, Shuangge 2006: Multiple augmentation with partial missing regressors. Biomedical Journal 48: 83–92.

Ma, Yanyuan, Marc G. Genton, Anastasios A. Tsiatis 2005: Locally efficient semiparametric estimators for generalized skew-elliptical distributions. Journal of the American Statistical Association 100: 980–989.

Maasoumi, E. 1990: How to live with misspecification if you must. Journal of Econometrics 44: 67–86.

Machamer, Peter, Lindley Darden, Carl F. Craver 2000: Thinking about mechanisms. Philosophy of Science 67: 1–25.

MacLehose, Richard F., Sol Kaufman, Jay S. Kaufman, Charles Poole 2005: Bounding causal effects under uncontrolled confounding using counterfactuals. Epidemiology 16: 548–555.

Madow, William G., Ingram Olkin (eds) 1983: Incomplete Data in Sample Surveys. 3 vols. New York: Academic Press.

Magnus, Jan R., Andrey L. Vasnev 2007: Local sensitivity and diagnostic tests. Econometrics Journal 10: 166–192.

Maguluri, Gangaji, Cun-Hui Zhang 1994: Estimation in the mean residual life regression model. Journal of the Royal Statistical Society B 56: 477–489.

Maldonado, George, Sander Greenland 2002: Estimating causal effects (with discussion). International Journal of Epidemiology 31: 422–429.

Mandel, Micha 2007: Censoring and truncation—Highlighting the differences. American Statistician 61: 321–324.

Manski, Charles F. 1989: Anatomy of the selection problem. The Journal of Human Resources 24: 343–360.

Manski, Charles F. 1990: Nonparametric bounds on treatment effects. American Economic Review 80: 319–323.

Manski, Charles F. 1993: The selection problem in econometrics and statistics. Pp. 73–84 in: G.S. Maddala, C.R. Rao, H.D. Vinod (eds) 1993: Handbook of Statistics, Vol. 11, Amsterdam: Elsevier.

Manski, Charles F. 1997: The mixing problem in programme evaluation. Review of Economic Studies 64: 537–553.

Manski, Charles F. 2003: Partial Identifiability of Probability Distributions. Berlin: Springer.

Manski, Charles F. 2004: Statistical treatment rules for heterogeneous populations. Econometrica 72: 1221–1246.

Manski, Charles F. 2005: Partial identification with missing data: Concepts and findings. International Journal of Approximate Reasoning 39: 151–165.

Manski, Charles F. 2007: Partial identification of counterfactual choice probabilities. International Economic Review 48: 1393–1410.

Manski, Charles F., Daniel S. Nagin 1998: Bounding disagreements about treatment effects: A case study of sentencing and recidivism. Sociological Methodology 28: 99–137.

Manski, Charles F., Elie Tamer 2002: Inference on regressions with interval data on a regressor or outcome. Econometrica 70: 519–546.

Manton, Kenneth G., Anatoli I. Yashin 1997: Effects of unobserved and partially observed covariate processes on system failure: a review of models and estimation strategies. Statistical Science 12: 20–34.

Marini, M.M., Burt Singer 1988: Causality in the social sciences. Pp. 347–409 in: Clogg, C.C. (ed) 1988: Sociological Methodology 18. San Francisco: Jossey-Bass.

Mark, Steven D., Hormuzd A. Katki 2006: Specifying and implementing nonparametric and semiparametric survival estimators in two-stage (nested) cohort studies with missing case data. Journal of the American Statistical Association 101: 460–471.

Marshall, Guillermo, Rolando de la Cruz-Mesia, Anna E. Barón, James H. Rutledge, Gary O. Zerbe 2006: Non-linear random effects model for multivariate responses with missing data. Statistics in Medicine 25: 2817–2830.

Martensen, Edwin P., Wiebe R. Pestman, Anthonius de Boer, Svetlana V. Belitser, Olaf H. Klungel 2006: Instrumental variables. Applications and limitations. Epidemiology 17: 260–267.

Martinussen, Torben 1999: Cox regression with incomplete covariate measurements using the EM-algorithm. Scandinavian Journal of Statistics 26: 479–491.

Martinussen, Torben, Thomas H. Scheike 2006: Dynamic Regression Models for Survival Data. Berlin: Springer.

Matheron, Georges 1989: Estimating and Choosing. An Essay on Probability in Practice. Berlin: Springer.

Matsui, Shigeyuki 2005: Stratified analysis in randomized trials with noncompliance. Biometrics 61: 816–823.

Matsuura, Masaaki, Shinto Eguchi 2005: Modeling late entry bias in survival analysis. Biometrics 61: 559–566.

Matthews, Abigail G., Dianne M. Finkelstein, Rebecca A. Betensky 2005: Analysis of familial aggregation in the presence of varying family size. Applied Statistics 54: 847–862.

Mayer, Karl Ulrich, Paul B. Baltes (eds) 1996: Die Berliner Altersstudie. Berlin: Akademie-Verlag.

McCaffrey, Daniel F., Greg Ridgeway, Andrew R. Morral 2004: Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychological Methods 9: 403–425.

McCandless, Lawrence C., Paul Gustafson, Adrian Levy 2006: Bayesian sensitivity analysis for unmeasured confounding in observational studies. Statistics in Medicine: DOI: 10.1002/sim.2711.

McCrorie, J. Roderick, Marcus J. Chambers 2006: Granger causality and the sampling of economic processes. Journal of Econometrics 132: 311–336.

McCullagh, Peter 2002: What is a statistical model? (with discussion). Annals of Statistics 30: 1225–1310.

McCullagh, Peter 2005: Exchangeability and regression models. Pp. 89–113 in: Davison, Anthony C., Yadolah Dodge, Nanny Wermuth (eds) 2005: Celebrating Statistics. Papers in Honour of Sir David Cox on his 80th Birthday. Oxford: Oxford University Press.

McCullagh, Peter 2008: Sampling bias and logistic models (with discussion). Journal of the Royal Statistical Society B 70: xxx–xxx.

McKim, V.R., S.P. Turner (eds) 1997: Causality in Crisis? Statistical Methods and the Search for Causal Knowledge in the Social Sciences. Notre Dame: University of Notre Dame Press.

McLeish, Donald L., Cyntha A. Struthers 2006: Estimation of regression parameters in missing data problems. Canadian Journal of Statistics 34: 233–259.

Mealli, Fabrizia, Donald B. Rubin 2002: Assumptions when analyzing randomized experiments with noncompliance and missing outcomes. Health Services& Outcomes Research Methodology 3: 225–232.

Meeden, Glen 2000: A decision theoretic approach to imputation in finite population sampling. Journal of the American Statistical Association 95: 586–595.

Meijer, Erik, Tom Wansbeek 2007: The sample selection model from a method of moments perspective. Econometric Review 26: 25–51.

Mellor, D.H. 1995: The Facts of Causation. London: Routledge.

Mengersen, K., S.A. Moynihan, R.L. Tweedie 2007: Causality and association: The statistical and legal approaches. Statistical Science 22: 227–254.

Menzies, Peter 2004: Causal models, token causation, and processes. Philosophy of Science 71: 820–832.

Mercatanti, Andrea 2004: Analyzing a randomized experiment with imperfect compliance and ignorable conditions for missing data: Theoretical and computational issues. Computational Statistics & Data Analysis 46: 493–509.

Mi, Jie 2006: MLE of parameters of location-scale distribution for complete and partially grouped data. Journal of Statistical Planning and Inference 136: 3567–3582.

Miller, Ruppert G. 1976: Least squares regression with censored data. Biometrika 63: 449–464.

Miller, Ruppert G. 1983: What price Kaplan–Meier? Biometrics 39: 1077–1082.

Miloslavsky, Maja, Sündüz Keleş, Mark J. van der Laan 2004: Recurrent events analysis in the presence of time-dependent covariates and dependent censoring. Journal of the Royal Statistical Society B 66: 239–257.

Minini, Pascal, Michael Chavance 2004a: Sensitivity analysis of longitudinal normal data with drop-outs. Statistics in Medicine 23: 1039–1054.

Minini, Pascal, Michael Chavance 2004b: Sensitivity analysis of longitudinal binary data with non-monotone missing values. Biostatistics 5: 531–544.

Miranda, Enrique, Inés Couso, Pedro Gil 2005: Random intervals as a model for imprecise information. Fuzzy Sets and Systems 154: 386–412.

Mitra, Nandita, Daniel F. Heitjan 2007: Sensitivity of the hazard ratio to non-ignorable treatment assignment in an observational study. Statistics in Medicine 26: 1398–1414.

Mnatsakanov, Robert, Frits H. Ruymgaart 2005: Some results for moment-empirical distribution functions. Nonparametric Statistics 17: 733–744.

Modarres, Reza 2003: Estimation of a bivariate symmetric distribution function. Statistics & Probability Letters 63: 25–34.

Mohler, Peter, Achim Koch, Siegfried Gabler 2003: Alles Zufall oder? Ein Diskussionsbeitrag zur Qualität von face to face Umfragen in Deutschland. ZUMA Nachrichten53: 10–15.

Mojirsheibani, Majid 2001: The Glivenko-Cantelli theorem based on data with randomly imputed missing values. Statistics & Probability Letters 55: 385–396.

Mojirsheibani, Majid 2007: Nonparametric curve estimation with missing data: A general empirical process approach. Journal of Statistical Planning and Inference 137: 2733–2758.

Mojirsheibani, Majid, Zahra Montazeri 2007: Statistical classification with missing covariates. Journal of the Royal Statistical Society B 69: 839–857.

Molchanov, Ilya 2005: Theory of Random Sets. Berlin: Springer.

Molenberghs, Geert, Caroline Beunckens, Cristina Sotto, Michael G. Kenward 2008: Every missing not at random model has a missingness at random counterpart with equal fit. Journal of the Royal Statistical Society B 70: 371–388..

Molenberghs, Geert, Herbert Thijs, Ivy Jansen, Caroline Beunckens, Michael G. Kenward, Craig Mallinckrodt, Raymond J. Carroll 2004: Analyzing incomplete longitudinal clinical trial data. Biostatistics 5: 445–464.

Molenberghs, Geert, Herbert Thijs, Michael G. Kenward, Geert Verbeke 2003: Sensitivity analysis of continuous incomplete longitudinal outcomes. Statistica Neerlandica 57: 112–135.

Molinaro, Annette M., Sandrine Dudoit, Mark J. van der Laan 2004: Tree-based multivariate regression and density estimation with right-censored data. Journal of Multivariate Analysis 90: 154–177.

Moodie, Erica E.M., Thomas S. Richardson, David A. Stephens 2007: Demystifying optimal dynamic treatment regimes. Biometrics 63: 447–455.

Morgan, Mary S. 1990: The history of Econometric Ideas. Cambridge: Cambridge University Press.

Morgan, T.M. 1986: Omitting covariates from the proportional hazards model. Biometrics 42: 993–995

Morgan, Stephen L., David J. Harding 2006: Matching estimators of causal effects. Prospects and pitfalls in theory and practice. Sociological Methods & Research 35: 3–60.

Morgenstern, Oskar 1963^2: On the Accuracy of Economic Observations. Princeton: Princeton University Press.

Morris, Ben J. 2002: Improved bounds for sampling contingency tables. Random Structures and Algorithms 21: 135–146.

Müller, Ursula U. 2007: Weighted least squares estimators in possibly misspecified nonlinear regression. Metrika 66: 39–59.

Müller, Ursula U., Anton Schick, Wolfgang Wefelmeyer 2008: Optimality of estimators for misspecified semi-Markov models. Stochastics 80: 181–196.

Mullin, Charles H. 2006: Identification and estimation with contaminated data: When do covariate data sharpen inference? Journal of Econometrics 130: 253–272.

Münnich, Ralf, Susanne Rässler 2005: PRIMA: A new multiple imputation procedure for binary variables. Journal of Official Statistics 21: 325–341.

Murnane, Richard J., Stuart Newstead, Randall J. Olsen 1985: Comparing public and private schools: The puzzling role of selectivity bias. Journal of Business & Economic Statistics 3: 23–35.

Murphy, Susan A. 2003: Optimal dynamic treatment regimes (with discussion). Journal of the Royal Statistical Society B 65: 331–366.

Murphy, Susan A., Bing Li 1995: Projected partial likelihood and its application to longitudinal data. Biometrika 82: 399–406.

Nakamura, Alice, Masao Nakamura 1998: Model specification and endogeneity. Journal of Econometrics 83: 213–237.

Nan, Bin, Mary J. Emond, Jon A. Wellner 2004: Information bounds for Cox regression models with missing data. Annals of Statistics 32: 723–753.

Nan, Bin, Menggang Yu, John D. Kalbfleisch 2006: Censored linear regression for case-cohort studies. Biometrika 93: 747–762.

Nandram, Balgobin, Jai Won Choi 2004: Nonparametric Bayesian analysis of a proportion for a small area under nonignorable nonresponse. Nonparametric Statistics 16: 821–839.

Neller, Katja 2005: Kooperation und Verweigerung: Eine Non-Response-Studie. ZUMA Nachrichten 57: 9–36.

Nelson, Barry L. 2002: Stochastic Modeling. Analysis & Simulation. Mineola: Dover.

Neugebauer, Romain, Mark van der Laan 2005: Why prefer double robust estimators in causal inference? Journal of Statistical Planning and Inference 129: 405–426.

Neugebauer, Romain, Mark van der Laan 2006a: Causal effects in longitudinal studies: Definition and maximum likelihood estimation. Computational Statistics & Sata Analysis 51: 1664–1675.

Neugebauer, Romain, Mark van der Laan 2006b: G-computation estimation for causal inference with complex longitudinal data. Computational Statistics & Sata Analysis 51: 1676–1697.

Neugebauer, Romain, Mark van der Laan 2007: Nonparametric causal effects based on marginal structural models. Journal of Statistical Planning and Inference 137: 419–434.

Neugebauer, Romain, Marshall M. Joffe, Ira B. Trager, Mark van der Laan 2007: Causal inference in longitudinal studies with history-restricted marginal structural models. Electronic Journal of Statistics 1: 119–154.

Neuhaus, John M., W.W. Hauck, John D. Kalbfleisch 1992: The effects of mixture distribution misspecification when fitting mixed-effects logistic models. Biometrika 79: 755–762.

Neyman, Jerzy 1960: Indeterminism in science and new demands on statisticians. Journal of the American Statistical Association 55: 625–639.

Newey, Whitney K., James L. Powell 1993: Efficiency bounds for some semiparametric selection models. Journal of Econometrics 58: 169–184.

Nguyen, Hung T. 2006: An Introduction to Random Sets. London: Chapman & Hall.

Nguyen, Hung T., B. Bouchon-Meunier 2003: Random sets and large deviations principle as a foundation for possibility measures. Soft Computing 8: 61–70.

Nguyen, Hung T., Berlin Wu 2006: Random and fuzzy sets in coarse data analysis. Comoutational Statistics & Data Analysis 51: 70–85.

Ní Bhrolcháin, Máire, Tim Dyson 2007: On causation in demography: Issues and illustrations. Population and Development Review 33: 1–36.

Nicoletti, Cheti 2002: Correcting for sample selection bias: Alternative estimators compared. http://www.iser.essex.ac.uk/activities/seminars/Monday_Afternoons/archive/papers/poverty3.pdf

Nicoletti, Cheti 2006: Nonresponse in dynamic panel data models. Journal of Econometrics 132: 461–489.

Nicoletti, Cheti, Franco Peracchi 2005: Survey response and survey characteristics: Microlevel evidence from the European Community Household Panel. Journal of the Royal Statistical Society A 168: 763–781.

Nicoletti, Cheti, Franco Peracchi 2006: The effect of income imputation on microanalyses: Evidence from the European Community Household Panel. Journal of the Royal Statistical Society A 169: 625–648.

Nielsen, Søren Feodor 1997: Inference and missing data: Asymptotic results. Scandinavian Journal of Statistics 24: 261–274.

Nielsen, Søren Feodor 2000: Relative coarsening at random. Statistica Neerlandica 54: 79–99.

Nielsen, Søren Feodor 2001: Nonparametric conditional mean imputation. Journal of Statistical Planning and Inference 99: 129–150.

Nielsen, Søren Feodor 2003a: Proper and improper multiple imputation (with discussion). International Statistical Review 71: 593–627.

Nielsen, Søren Feodor 2003b: Survival analysis with coarsely observed covariates. SORT 27: 41–64.

Nielsen, Søren Feodor 2005: Local linear estimating equations: Uniform consistency and rate of convergence. Journal of Nonparametric Statistics 17: 493–513.

Nishii, R. 1988: Maximum likelihood principle and model selection when the true model is unspecified. Journal of Multivariate Analysis 27: 392–403.

Nitsch, D., B.L. Stavola, S.M.B. Morton, D.A. Leon 2006: Linkage bias in estimating the association between childhood exposures and propensity to become a mother: An example of simple sensitivity analyses. Journal of the Royal Statistical Society A 169: 493–505.

Nittner, Thomas 2003: Missing at random in nonparametric regression–A simulation experiment. Statistical Methods & Applications 12: 195–210.

Nordholt, Eric Schulte 1998: Imputation: Methods, simulation experiments and practical examples. International Statistical Review 66: 157–180.

Norkus, Zenonas 2005: Mechanisms as miracle makers? The rise and inconsistencies of the "mechanismic approach" in social science and history. History and Theory 44: 348–372.

Oakes, David, John Ritz 2000: Regression in a bivariate copula model. Biometrika 87: 345–352.

Oakes, David, Antai Wang 2003: Copula model generated by Dabrowka's association measure. Biometrika 90: 478–481.

O'Brien, Peter C., David Zhang, Kent R. Bailey 2005: Semi-parametric and non-parametric methods for clinical trials with incomplete data. Statistics in Medicine 24: 341–358.

O'Hagan, Anthony 2003: HSSS model criticism (with discussion). Pp. 423–453 in: Peter J. Green, Nils Lid Hjort, Sylvia Richardson (eds.) 2003: Highly Structured Stochastic Systems. Oxford: Oxford University Press.

O'Hara Hines, R.J., W.G.S. Hines 2005: An appraisal of methods for the analysis of longitudinal categorical data with MAR drop-outs. Statistics in Medicine 24: 3549–3563.

Olesen, Anne Vingaard, Erik Thorlund Parner 2006: Correcting for selection using frailty models. Statistics in Medicine 25: 1672–1684.

Oller, Ramon, Guadalupe Gómez, M. Luz Calle 2004: Interval censoring: Model characterizations for the validity of the simplified likelihood. Canadian Journal of Statistics 32: 315–326.

Oller, Ramon, Guadalupe Gómez, M. Luz Calle 2007: Interval censoring: Identifiability and the constant-sum property. Biometrika 94: 61–70.

O'Malley, A. James, Sharon-Lise T. Normand 2005: Likelihood methods for treatment noncompliance and subsequent nonresponse in randomized trials. Biometrics 61: 325–334.

Opp, Karl-Dieter 2005: Explanations by mechanisms in the social sciences. Problems, advantages and alternatives. Mind & Society 4: 163–178.

Orchard, Terence, Max A. Woodbury 1972: A missing information principle: theory and applications. Pp. 697–715 in: Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, New York: Prentice-Hall.

Paddock, Susan M. 2002: Bayesian nonparametric multiple imputation of partially observed data with ignorable non-response. Biometrika 89: 529–538.

Paik, Myunghee Cho 1997a: The generalized estimating equation approach when data are not missing completely at random. Journal of the American Statistical Association 92: 1320–1329.

Paik, Myunghee Cho 1997b: Multiple imputation for the Cox proportional hazards model with missing covariates. Lifetime Data Analysis 3: 289–298.

Paik, Myunghee Cho 2004: Nonignorable missingness in matched case-control data analyses. Biometrics 60: 306–314.

Paliwal, Prashni, Alan E. Gelfand 2006: Estimating measures of diagnostic accuracy when some covariate information is missing. Statistics in Medicine 25: 2981–2993.

Pan, W., Charles Kooperberg 1999: Linear regression for bivariate censored data via multiple imputation. Statistics in Medicine 18: 3111–3121.

Pantazis, N., G. Touloumi, A.S. Walker, A.G. Babiker 2005: Bivariate modelling of longitudinal measurements of two human immunodeficiency type 1 disease progression markers in the presence of informative drop-outs. Applied Statistics 54: 405–423.

Park, Soomin, Mari Palta, Jun Shao, Lei Shen 2002: Bias adjustment in analysing longitudinal data with informative missingness. Statistics in Medicine 21: 277–291.

Park, Yuhyun, Lu Tian, L.J. Wei 2006: One- and two-sample nonparametric inference procedures in the presence of a mixture of independent and dependent censoring. Biostatistics 7: 252–267.

Parner, J., Elja Arjas 1999: Causal reasoning from longitudinal data. Research Reports A27, Rolf Nevanlinna Institute, Helsinki. Appeared as E. Arjas, J. Parner: Causal reasoning from longitudinal data. Scandinavian Journal of Statistics 31 (2004): 171–201 (with a discussion by S.L. Lauritzen and O.O. Aalen and replies by D.B. Rubin and E. Arjas)

Parzen, Michael, Stuart R. Lipsitz 2007: Perturbing the minimand resampling with Gamma(1,1) random variables as an extension of the Bayesian bootstrap. Statistics & Probability Letters 77: 654–657.

Parzen, Michael, Stuart R. Lipsitz, Garrett M. Fitzmaurice 2005: A note on reducing the bias of the approximate Bayesian bootstrap imputation variance estimator. Biometrika 92: 971–974.

Parzen, Michael, Stuart R. Lipsitz, Garrett M. Fitzmaurice, Joseph G. Ibrahim, Andrea Troxel 2006: Pseudo-likelihood methods for longitudinal binary data with non-ignorable missing responses and covariates. Statistics in Medicine 25: 2784–2796.

Patilea, Valentin 2001: Convex models, MLE and misspecification. Annals of Statistics 29: 94–123.

Patilea, Valentin, Jean-Marie Rolin 2006: Product-limit estimators of the survival function with twice censored data. Annals of Statistics 34: 925–938.

Pearl, Judea 1999: Probabilities of causation: Three counterfactual interpretations and their identification. Synthese 121: 93–149.

Pearl, Judea 2000: Causality: Models, Reasoning, and Inference. Cambridge: Cambridge University Press.

Pearl, Judea 2003: Statistics and causal inference: A review (with discussion). Test 12: 281–345.

Pearl, Judea, T.S. Verma 1992: A statistical semantics for causation. Statistics & Computing 2: 91–95.

Peng, Limin, Jason P. Fine 2006: Rank estimation of accelerated lifetime models with dependent censoring. Journal of the American Statistical Association 101: 1085–1093.

Peng, Limin, Jason P. Fine 2007: Regression modeling of semicompeting risks data. Biometrics 63: 96–108.

Peng, Yahong, Roderick J.A. Little, Trivellore E. Raghunathan 2004: An extended general location model for causal inference from data subject to noncompliance and missing values. Biometrics 60: 598–607.

Persson, Johannes 2006: Compartment causation. Synthese 149: 535–550.

Petersen, Trond 1991: Time aggregation bias in continuous-time hazard-rate models. pp. 261–290 in: Marsden, P.V. (ed) 1991: Sociological Methodology 21. Oxford: Blackwell.

Petersen, Trond, K.W. Koput 1992: Time-aggregation bias in hazard-rate models with covariates. Sociological Methods & Research 21: 25–51.

Petersen, Maya L., Sandra E. Sinisi, Mark J. van der Laan 2006: Estimation of direct causal effects. Epidemiology 17: 276–284.

Pfanzagl, Johann 1990: Estimation in Semiparametric Models. Some Recent Developments. Berlin: Springer.

Pfanzagl, Johann 1994: Parametric Statistical Theory. Berlin: Walter de Gruyter.

Pfanzagl, Johann 1998: The nonexistence of confidence sets for discontinuous functionals. Journal of Statistical Planning and Inference 75: 9–20.

Phelps, Robert R. 2001^2: Lectures on Choquet's Theorem. Berlin: Springer.

Phillips, Carl V., Karen J. Goodman 2006: Causal criteria and counterfactuals; nothing more (or less) than scientific common sense. Emerging Themes in Epidemiology 3: 5.

Pietroski, Paul M. 2000: Causing Actions. Oxford: Oxford University Press.

Pierce, Donald A. 1982: The asymptotic effect of substituting estimators for parameters in certain types of statistics. Annals of Statistics 10: 475–478.

Pierce, Donald A., Dawn Peters 1999: Improving on exact tests by approximate conditioning. Biometrika 86: 265–277.

Pistone, Giovanni, Eva Riccomagno, Henry P. Wynn 2001: Algebraic Statistics. Computational Commutative Algebra in Statistics. London: Chapman & Hall.

Pitman, E.J.G. 1979: Some Basic Theory of Statistical Inference. London: Chapman & Hall.

Poirier, Dale J., Paul A. Ruud 1981: On the appropriateness of endogenous switching. Journal of Econometrics 16: 249–256.

Pollard, David 1984: Convergence of Stochastic Processses. Berlin: Springer.

Pollard, David 2002: A User's Guide to Measure Theoretic Probability. Cambridge: Cambridge University Press.

Pons, Odile 2002: Estimation in the Cox model with missing covariate data. Nonparametric Statistics 14: 223–247.

Pons, Odile 2006: Estimation for semi-Markov models with partial observations via self-consistency equations. Statistics 40: 377–388.

Potthoff, Richard F., Gail E. Tudor, Karen S. Pieper, Vic Hasselblad 2006: Can one assess whether missing data are missing at random in medical studies? Statistical Methods in Medical Research 15: 213–234.

Pratt, J.W., R. Schlaifer 1984: On the nature and discovery of structure. Journal of the American Statistical Association 79: 9–33.

Prentice, Ross L. 1999: On non-parametric maximum likelihood estimation of the bivariate survivor function. Statistics in Medicine 18: 2517–2527.

Prentice, Ross L., J. Cai 1992: Covariance and survivor function estimation using censored multivariate failure time data. Biometrika 79:495–512.

Prentice, Ross L., F. Zoe Moodie, Jianrong Wu 2004: Hazard based nonparametric survivor function estimation. Journal of the Royal Statistical Society B 66: 305–319.

Prentice, Ross L., Mary Pettinger, Garnet L. Anderson 2005: Statistical issues arising in the Women's Health Initiative (with discussion). Biometrics 61: 899–941.

Pruitt, R.C. 1991: On negative mass assigned by the bivariate Kaplan–Meier estimator. Annals of Statistics 19: 443–453.

Pruitt, R.C. 1993: Small sample comparison of six bivariate survival curve estimators. Journal of Statistical Computation and Simulation 45: 147–167.

Qi, Lihong, C.Y. Wang, Ross L. Prentice 2005: Weighted estimators for proportional hazards regression with missing covariates. Journal of the American Statistical Association 100: 1250–1263.

Qin, Jing, Denis Leung, Jun Shao 2002: Estimation with survey data under nonignorable nonresponse or informative sampling. Journal of the American Statistical Association 97: 193–200.

Qin, Jing, Biao Zhang 2007: Empirical-likelihood-based inference in missing response problems and its application in observational studies. Journal of the Royal Statistical Society B 69: 101–122.

Qin, Jing, Biao Zhang 2008: Empirical-likelihood-based difference-in-differences estimators. Journal of the Royal Statistical Society B 70: 329–349.

Qu, Annie, Peter X.-K. Song 2002: Testing ignorable missingness in estimating equation approaches for longitudinal data. Biometrika 89: 841–850.

Ramsey, Frank P. 1990: Philosophical Papers. Edited by D.H. Mellor. Cambridge: Cambridge University Press.

Rao, J.N.K. 1993: Jackknife variance estimation with imputed survey data. Proceedings of the Survey Research Methods Section, American Statistical Association 31–40.

Rao, J.N.K. 1996: On variance estimation with imputed survey data. Journal of the American Statistical Association 91: 499–506.

Rao, M.M. 2005^2: Conditional Measures and Applications. London: Chapman & Hall.

Rässler, Susanne 2002: Statistical Matching. A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches. Berlin: Springer.

Rässler, Susanne, Regina T. Riphahn 2006: Survey item nonresponse and its treatment. Allgemeines Statistisches Archiv 90: 217–232.

Rathbun, Stephen L., Saul Shiffman, Chad J. Gwaltney 2007: Modelling the effects of partially observed covariates on Poisson process intensity. Biometrika 94: 153–165.

Rathouz, Paul J. 2003: Likelihood methods for missing covariate data in highly stratified studies. Journal of the Royal Statistical Society B 65: 711–723.

Rathouz, Paul J. 2004: Fixed effect models for longitudinal binary data with drop-outs missing at random. Statistica Sinica 14: 969–988.

Rathouz, Paul J. 2007: Identifiability assumptions for missing covariate data in failure time regression models. Biostatistics 8: 345–356.

Reilly, Marie, Margaret Pepe 1997: The relationship between hot-deck multiple imputation and weighted likelihood. Statistics in Medicine 16: 5–19.

Reiss, Rolf-Dieter 1993: A Course on Point Processes. Berlin: Springer.

Reiter, Jerome O., Trivellore E. Raghunathan 2007: The multiple adaptations of multiple imputation. Journal of the American Statistical Association 102: 1462–1471.

Ren, Jian-Jian 2003: Regression *M*-estimators with non-iid doubly censored data. Annals of Statistics 31: 1186–1219.

Rendtel, Ulrich 1995: Panelausfälle und Panelrepräsentativität. Frankfurt: Campus.

Richardson, Thomas S., Peter Spirtes 2002: Ancestral graph Markov models. Annals of Statistics 30: 962–1030.

Richardson, Thomas S., Peter Spirtes 2003: Causal inference via ancestral graph models (with discussion). Pp. 83–113 in: Peter J. Green, Nils Lid Hjort, Sylvia Richardson (eds.) 2003: Highly Structured Stochastic Systems. Oxford: Oxford University Press.

Rieder, Helmut 1994: Robust Asymptotic Statistics. Berlin: Springer.

Riphahn, Regina T., Oliver Serfling 2002: Item non-response on income and wealth questions. Forschungsinstitut zur Zukunft der Arbeit. Discussion Paper No. 573:

Ritov, Ya'acov 1990: Estimation in a linear regression model with censored data. Annals of Statistics 18:303–328.

Rivest, Louis-Paul, Martin T. Wells 2001: A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. Journal of Multvariate Analysis 79: 138–155.

Robins, James M. 1997: Non-response models for the analysis of non-monotone non-ignorable missing data. Statistics in Medicine 16: 21–37.

Robins, James M. 1999: Association, causation, and marginal structural models. Synthese 121: 151–179.

Robins, James M., Richard D. Gill 1997: Non-response models for the analysis of non-monotone ignorable missing data. Statistics in Medicine 16: 39–56.

Robins, James M., Ya'acov Ritov 1997: Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. Statistics in Medicine 16: 285–319.

Robins, James M., Andrea Rotnitzky, Lue Ping Zhao 1994: Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association 89: 846– 866.

Robins, James M., Andrea Rotnitzky 1995: Semiparametric efficiency in multivariate regression models with missing data. Journal of the American Statistical Association 90: 122–129.

Robins, James M., Andrea Rotnitzky, Lue Ping Zhao 1995: Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association 90: 106–121.

Robins, James M., Andrea Rotnitzky, Daniel O. Scharfstein 2000: Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: M.E. Halloran, D. Berry (eds) 2000: Statistical Models in Epidemiology: The Environment and Clinical Trials. Berlin: Springer.

Robins, James M., Andrea Rotnitzky 2004: Estimation of treatment effects in randomised trials with non-compliance and a dichotomous outcome using structural mean models. Biometrika 91: 763–783.

Robins, James M., Richard Scheines, Peter Spirtes, Larry Wasserman 2003: Uniform consistency in causal inference. Biometrika 90: 491–515.

Robins, James M., Anastasios A. Tsiatis 1992: Semiparametric estimation of an accelerated failure time model with time-dependent covariates. Biometrika 79: 311–319.

Robins, James M., Naisyin Wang 2000: Inference for imputation estimators. Biometrika 87: 113–124.

Robins, James M., Larry Wasserman 1999: On the impossibility of inferring causation from association without background knowledge (with discussion). Pp. 305–345 in: Glymour, Clark, Gregory F. Cooper (eds) 1999: Computation, Causation, and Discovery. Menlo Park: AAAI Press.

Robins, James M., Larry Wasserman 2000: Conditioning, likelihood, and coherence: A review of some foundational concepts. Journal of the American Statistical Association 95: 1340–1346.

Rohwer, Götz, Ulrich Pötter 2001a: Grundzüge der sozialwissenschaftlichen Statistik. Weinheim: Juventa.

Rohwer, Götz, Ulrich Pötter 2001b: Kausale und funktionale Erklärungen in der Sozialforschung. http://www.stat.rub.de/papers/dkfe.ps

Rohwer, Götz, Ulrich Pötter 2002a: Methoden sozialwissenschaftlicher Datenkonstruktion. Weinheim: Juventa.

Rohwer, Götz, Ulrich Pötter 2002b: Wahrscheinlichkeit. Begriff und Rhetorik in der Sozialforschung. Weinheim: Juventa.

Rojas, J. Maurice 2003: Why polyhedra matter in non-linear equation solving. Contemporary Mathematics 334: 293–320.

Romano, Joseph P. 2004: On non-parametric testing, the uniform behaviour of the t-test, and related problems. Scandinavian Journal of Statistics 31: 567–584.

Rosenbaum, Paul R. 2002^2: Observational Studies. Berlin: Springer. First edition 1995.

Rosenbaum, Paul R. 2003: Exact confidence intervals for nonconstant effects by inverting the signed rank test. The American Statistician 57: 132–138.

Rosenbaum, Paul R. 2005: Sensitivity analysis in observational studies. Pp. 1809–1814 in: Brian S. Everitt, David C. Howell (eds) 2005: Encyclopedia of Statistics in Behavioral Science. Vol. 4. New York: Wiley.

Rosenbaum, Paul R., Jeffrey H. Silber 2001: Matching and thick description in an observational study of mortality after surgery. Biostatistics 2: 217–232.

Rotnitzky, Andrea, David R. Cox, Matteo Bottai, James M. Robins 2000: Likelihood-based inference with singular information matrix. Bernoulli 6: 243–284.

Rotnitzky, Andrea, Andres Farall, Andrea Bergesio, Daniel Scharfstein 2007: Analysis of failure time data under competing censoring mechanisms. Journal of the Royal Statistical Society B 69: 307–327.

Rotnitzky, Andrea, James M. Robins 1997: Analysis of semi-parametric regression models with non-ignorable non-response. Statistics in Medicine 16: 81–102.

Rotnitzky, Andrea, James M. Robins, Daniel O. Scharfstein 1998: Semiparametric regression for repeated outcomes with nonignorable nonresponse. Journal of the American Statistical Association 93: 1321–1339.

Roussas, George G. 1972: Contiguity of Probability Measures. Some Applications in Statistics. Cambridge. Cambridge University Press.

Roy, Jason 2003: Modeling longitudinal data with nonignorable dropouts using a latent dropout class model. Biometrics 59: 829–836.

Roy, Jason 2007: Latent class models and their application to missing-data patterns in longitudinal studies. Statistical Methods in Medical Research 16: 441–456.

Roy, Jason, Xihong Lin 2005: Missing covariates in longitudinal data with informative dropouts: Bias analysis and inference. Biometrics 61: 837–846.q

Roy, Jason, Don Alderson, Joseph W. Hogan, Karen T. Tashima 2006: Conditional inference methods for incomplete Poisson data with endogeneous time-varying covariates: Emergency department use among HIV-infected women. Journal of the American Statistical Association 101: 424–434.

Royall, Richard 1997: Statistical Evidence. A Likelihood Paradigm. London: Chapman & Hall.

Royall, Richard, T.-S. Tsou 2003: Interpreting statistical evidence by using imperfect models: robust adjusted likelihood functions. Journal of the Royal Statistical Society B 65: 391–404.

Ruan, Ping K., Robert J. Gray 2006: A method for analyzing disease-specific mortality with missing cause of death information. Lifetime Data Analysis 12: 35–51.

Rubin, Daniel, Mark J. van der Laan 2005: A general imputation methodology for nonparametric regression with censored data. U.C. Berkeley Division of Biostatistics Working Paper Series 194.

Rubin, Daniel, Mark J. van der Laan 2007: A doubly robust censoring unbiased transformations. The International Journal of Biostatistics 3: Article 4.

Rubin, Donald B. 1974: Characterizing the estimation of parameters in incomplete-data problems. Journal of the American Statistical Association 69: 467–474.

Rubin, Donald B. 1976: Inference and missing data (with discussion). Biometrika 63:581–592.

Rubin, Donald B. 1977: Formalizing subjective notions about the effect of nonrespondents in sample surveys. Journal of the American Statistical Association 72: 538–543.

Rubin, Donald B. 1991: Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. Biometrics 47: 1213–1234.

Rubin, Donald B. 1996: Multiple imputation after 18+ years. Journal of the American Statistical Association 91: 473–489.

Rubin, Donald B. 2003: Nested multiple imputation of NMES via partially incompatible MCMC. Statistica Neerlandica 57: 3–18.

Rubin, Donald B. 2004: Direct and indirect causal effects via potential outcomes. Scandinavian Journal of Statistics 31: 161–170.

Rubin, Donald B. 2006: Causal inference through potential outcomes and principal stratification: Application to studies with "censoring" due to death (with discussion). Statistical Science 21: 299–321.

Rubin, Donald B. 2007: The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. Statistics in Medicine 26: 20–36.

Rubin, Donald B., Elizabeth A. Stuart 2006: Affinely invariant matching methods with discriminant mixtures of proportional ellipsoidally symmetric distributions. Annals of Statistics 34: 1814–1826.

Rubin, Donald B., Neal Thomas 2000: Combining propensity score matching with additional adjustments for prognostic covariates. Journal of the American Statistical Association 95: 573–585.

Rubin, Donald B., Richard P. Waterman 2006: Estimating the causal effects of marketing interventions using propensity score methodology. Statistical Sience 21: 206–222.

Rubin-Bleuer, Susana, Ioana Schiopu Kratina 2005: On the two-phase framework for joint model and design-based inference. Annals of Statistics 33: 2789–2810.

Rueda, M., S. González, A. Arcos 2007: A predictive estimator of the mean with missing data. Quality & Quantity 41: 201–217.

Rueda, M., S. Martínez, H. Martínez, A. Arcos 2006: Mean estimation with calibration techniques in presence of missing data. Computational Statistics & Data Analysis 50: 3263–3277.

Russo, Federica 2006: Salmon and van Fraassen on the existence of unobservable entities: A matter of interpretation of probability. Foundations of Science 11: 221–247.

Ruud, P.A. 1986: Consistent estimation of limited dependent variable models despite misspecification of distribution. Journal of Econometrics 32: 157–187.

Saha, Chandan, Michael P. Jones 2005: Asymptotic bias in the linear mixed effects model under non-ignorable missing data mechanisms. Journal of the Royal Statistical Society B 67: 167–182.

San Martín, Ernesto 2005: Ignorable common information, null sets and Basu's first theorem. Sankhyā 67: 647–698.

Särndal, Carl-Erik, Bengt Swensson, Jan Wretman 1992: Model Assisted Survey Sampling. Berlin: Springer.

Sartori, Nicola, Alberto Salvan, Karl Thomaseth 2005: Multiple imputation of missing values in a cancer mortality analysis with estimated exposure dose. Computational Statistics & Data Analysis 49: 937–953.

Satten, Glen A., Sonmath Datta 2001: The Kaplan-Meier estimator as an inverse-probability-of-censoring weighted average. The American Statistician 55: 207–210.

Savage, Leonard J. 1972: The Foundations of Statistics. New York: Dover.

Schafer, Joseph L. 1997: Analysis of Incomplete Multivariate Data. London: Chapman & Hall.

Schafer, Joseph L. 2003: Multiple imputation in multivariate problems when the imputation and analysis models differ. Statistica Neerlandica 57: 19–35.

Schafer, Joseph L., John W. Graham 2002: Missing data: Our view of the state of the art. Psychological Methods 7: 147–177.

Schafer, Joseph L., Nathaniel Schenker 2000: Inference with imputed conditional means. Journal of the American Statistical Association 95: 144–154.

Schaffner, Julie Anderson 1998: Generating conditional expectations from models with selectivity bias: Comment. Economics Letters 58: 255–261.

Schafgans, Marcia M.A. 2004: Finite sample properties for the semiparametric estimation of the intercept of a censored regression model. Statistica Neerlandica 58: 35–56.

Scharfstein, Daniel O., Andrea Rotnitzky, James M. Robins 1999: Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). Journal of the American Statistical Association 94: 1096–1146.

Scharfstein, Daniel O., Charles F. Manski, James C. Anthony 2004: On the construction of bounds in prospective studies with missing ordinal outcomes: Application to the Good Behavior Game Trial. Biometrics 60: 154–164.

Scharfstein, Daniel O., James M. Robins 2002: Estimation of the failure time distribution in the presence of informative censoring. Biometrika 89: 617–634.

Scharfstein, Daniel O., Michael J. Daniels, James M. Robins 2003: Incorporating prior beliefs about selection bias into the analysis of randomized trials with missing outcomes. Biostatistics 4: 495–512.

Scharfstein, Daniel O., Rafael A. Irizarry 2003: Generalized additive selection models for the analysis of studies with potentially nonignorable missing outcome data. Biometrics 59: 601–613.

Schaubel, Douglas, Jianwei Cai 2006a: Rate/mean regression for multiple-sequence recurrent event data with missing event category. Scandinavian Journal of Statistics 33: 191–207.

Schaubel, Douglas, Jianwei Cai 2006b: Multiple imputation methods for recurrent event data with missing event category. Canadian Journal of Statistics 34: 677–692.

Scheffler, U. 1994: Token versus type causation. Pp. 91–108 in: Faye, J., U. Scheffler, M. Urchs (eds) 1994: Logic and Causal Reasoning. Berlin: Akademie Verlag.

Scheike, Thomas H., Mei-Jie Zhang 2007: Direct modelling of regression effects for transition probabilities in multistate models. Scandinavian Journal of Statistics 34: 17–32.

Schenker, Nathaniel, Trivellore E. Raghunathan, Pei-Lu Chiu, Diane M. Makuc, Gangyu Zhang, Alan J. Cohen 2006: Multiple imputation of missing income data in the National Health Interview Survey. Journal of the American Statistical Association 101: 924–933.

Schenker, Nathaniel, A.H. Welsh 1988: Asymptotic results for multiple imputation. Annals of Statistics 16: 1550–1566.

Schepers, J., Gert Wagner 1989: Soziale Differenzen der Lebenserwartung in der Bundesrepublik Deutschland – Neue empirische Analysen. Zeitschrift für Sozialreform 35: 670–682.

Scheuren, Fritz 2005: Multiple imputation: How it began and continues. American Statistician 59: 315–319.

Schmee, J., G.J. Hahn 1979: A simple method for regression analysis with censored data. Technometrics 21: 417–432.

Schneekloth, Ulrich, Ingo Leven 2003: Woran bemisst sich eine „gute" allgemeine Bevölkerungsumfrage? Analysen zu Ausmaß, Bedeutung und zu den Hintergründen von Nonresponse in zufallsbasierten Stichprobenerhebungen am Beispiel des ALLBUS. ZUMA Nachrichten 53: 16–57.

Schräpler, Jörg-Peter 2004: Respondent behavior in panel studies — A case study for income-nonresponse by means of the SOEP. Sociological Methods & Research 33: 118–156.

Schreiber, T. 2000: Statistical inference from set-valued observations. Probability Mathematical Statistics 20: 223–235.

Schubnell, H. , Herberger, L. 1970. Die Volkszaehlung am 27. Mai 1970. Wirtschaft und Statistik 22, Heft 4, 179 – 185.

Schweder, T. 1970: Composable Markov processes. Journal of Applied Probability 7: 400–410.

Schweder, T. 1986: Kan sosialstatistikeren skille årsak fra virkning? Tidsskrift for Samfunnsforskning 27: 357–369.

Seidenfeld, Teddy, Larry Wasserman 1998: Dilation for sets of probabilities. Annals of Statistics 21: 1139–1154.

Senn, Stephen 2004: Added values. Controversies concerning randomization and additivity in clinical trials. Statistics in Medicine 23: 3729–3753.

Senn, Stephen, Erika Graf, Angelika Caputo 2007: Stratification for the propensity score compared with linear regression techniques to asses the effect of treatment or exposure. Statistics in Medicine 26: 5529–5544.

Severini, Thomas A. 2000: Likelihood Methods in Statistics. Oxford: Oxford University Press.

Severini, Thomas A., H. Wong 1992: Profile likelihood and conditionally parametric models. Annals of Statistics 20: 1768–1802.

Shafer, Glenn 1976: A Mathematical Theory of Evidence. Princeton: Princeton University Press.

Shafer, Glenn 1995: The situation of causality. Foundations of Science 1: 543–563.

Shaked, Moshe, J.G. Shanthikumar 1991: Dynamic multivariate mean residual life functions. Journal of Applied Probability 28: 613–629.

Shao, Jun, Hansheng Wang 2008: Confidence intervals based on survey data with nearest neighbor imputation. Statistica Sinica 18: 281–297.

Shardell, Michelle, Ram R. Miller 2008: Weighted estimating equations for longitudinal studies with death and non-monotone missing time-dependent covariates and outcomes. Statistics in Medicine 27: 1008–1025.

Shardell, Michelle, Daniel O. Scharfstein, Noya Galai, David Vlahov, Samuel A. Bozzette 2004: Sensitivity analysis for informatively interval-censored discrete time-to-event data. Dept. of Biostatistics Working Paper 36, Johns Hopkins University.

Shardell, Michelle, Daniel O. Scharfstein, Samuel A. Bozzette 2007: Survival curve estimation for informatively coarsened discrete event-time data. Statistics in Medicine 26: 2184–2202.

Shen, Changyu, Lisa Weissfeld 2005: Application of pattern-mixture models to outcomes that are potentially missing not at random using pseudo maximum likelihood estimation. Biostatistics 6: 333–347.

Shen, Changyu, Lisa Weissfeld 2006: A copula model for reapeated measurements with non-ignorable non-monotone missing outcome. Statistics in Medicine 25: 2427–2440.

Shen, Pao-sheng 2003: The product-limit estimate as an inverse-probability weighted average. Communications in Statistics. Theory and Methods 32: 1119–1133.

Shen, Pao-sheng 2006: An inverse-probability-weighted approach to estimation of the bivariate survival function under left-truncation and right-censoring. Journal of Statistical Planning and Inference 136: 4365–4384.

Sheng, Xiaoming, K.C. Carrière 2005: Strategies for analysing missing item response data with an application to lung cancer. Biometrical Journal 47: 605–615.

Shepherd, Bryan E., Peter B. Gilbert, Thomas Lumley 2007: Sensitivity analyses comparing time-to-event outcomes existing only in a subset selected postrandomization. Journal of the American Statistical Association 102: 573–582.

Shorack, Galen R. 2000: Probability for Statisticians. Berlin: Springer.

Siannis, Fotios 2004: Applications of a parametric model for informative censoring. Biometrics 60: 704–714

Siannis, Fotios, John Copas, Guobing Lu 2005: Sensitivity analysis for informative censoring in parametric survival models. Biostatistics 6: 77–91.

Silva, J.M.C. Santos 2003: A note on the estimation of mixture models under endogenous sampling. Econometrics Journal 6: 46–52.

Skinner, Chris J. 1987: Comment on "Parameter orthogonality and approximate conditional inference" by Cox and Reid. Journal of the Royal Statistical Society B 49: 24

Skinner, Chris J. 1989: GLM's and coefficient ratios. in: GLIM 89 Proceedings, Trento, Italy. Berlin: Springer

Skinner, Chris J., J.N.K. Rao 2002: Jackknife variance estimation for multivariate statistics under hot-deck imputation from common donors. Journal of Statistical Planning and Inference 102: 149–167.

Slud, E. 1992: Partial likelihood for continuous-time stochastic processes. Scandinavian Journal of Statistics 19: 97–109.

Small, Christopher G., D.L. McLeish 1994: Hilbert Space Methods in Probability and Statistical Inference. New York: Wiley.

Small, Dylan S., Joseph L. Gastwirth, Abba M. Krieger, Paul R. Rosenbaum 2006: *R*-estimates vs. GMM: A theoretical case study of validity and efficiency. Statistical Science 21: 363–375.

Small, Dylan S., Thomas R. Ten Have, Paul R. Rosenbaum 2008: Randomization inference in a group-randomized trial of treatments for depression: Covariate adjustment, noncompliance, and quantile effects. Journal of the American Statistical Association 103: 271–279.

Smith, Jeffrey A., Petra E. Todd 2001: Reconciling conflicting evidence on the performance of propensity-score matching methods. American Economic Review 91: 112–118.

Smith, Jeffrey A., Petra E. Todd 2005a: Does matching overcome LaLonde's critique of nonexperimental estimators? Journal of Econometrics 125: 305–353.

Smith, Jeffrey A., Petra E. Todd 2005b: Rejoinder. Journal of Econometrics 125: 365–375.

Smith, Murray D. 2003: Modelling sample selection using Archimedean copulas. Econometrics Journal 6: 99–123.

Sobel, M. 1997: Measurement, causation, and local independence in latent variable models. Pp. 11–28 in: M. Berkane (ed) 1997: Latent Variable Modeling and Applications to Causality. Berlin: Springer.

Solomon, P.J. 1984: Effect of misspecification of regression models in the analysis of survival data. Biometrika 71: 291–298.

Solomon, P.J. 1986: Effect of misspecification of regression models in the analysis of survival data. Correction. Biometrika 73: 245.

Song, Juwon, Thomas R. Belin 2004: Imputation for incomplete high-dimensional multivariate normal data using a common factor model. Statistics in Medicine 23: 2827–2843.

Song, Xiao, Shuangge Ma, Jian Huang, Xiao-Hua Zhou 2007: A semiparametric approach for the nonparametric transformation survival model with multiple covariates. Biostatistics 8: 197–211.

Sørensen, Aage B. 1998: Theoretical mechanisms and the empirical study of social processes. In: Hedström, Peter, Richard Swedberg (eds) 1998: Social Mechanisms. An Analytical Approach to Social Theory. Cambridge: Cambridge University Press.

Spiess, Martin 2006: Estimation of a two-equation panel model with mixed continuous and ordered categorical outcomes and missing data. Applied Statistics 55: 525–538.

Spirtes, Peter 2005: Graphical models, causal inference, and econometric models. Journal of Economic Methodology 12: 3–34.

Squire, Peverill 1988: Why the 1936 *Literary Digest* poll failed. Public Opinion Quarterly 52: 125–133.

Stadje, Wolfgang 2005: The evolution of aggregated Markov chains. Statistics & Probability Letters 74: 303–311.

Stanley, Richard P. 1997: Enumerative Combinatorics I. Cambridge: Cambridge University Press.

Stanley, Richard P. 1999: Enumerative Combinatorics II. Cambridge: Cambridge University Press.

Statistisches Bundesamt (ed), 2000: Datenreport 1999: Zahlen und Fakten über die Bundesrepublik Deutschland. Bundeszentrale für politische Bildung, Schriftenreihe Band 365: Bonn.

Statistisches Bundesamt (ed), 2004: Datenreport 2004: Zahlen und Fakten über die Bundesrepublik Deutschland. Bundeszentrale für politische Bildung, Schriftenreihe Band 450: Bonn.

Statistisches Bundesamt 2001: Statistisches Jahrbuch. Wiesbaden.

Steel, Daniel 2004: Social mechanisms and causal inference. Philosophy of the Social Sciences 34: 55–78.

Steel, Daniel 2005a: Indeterminism and the causal Markov condition. British Journal for the Philosophy of Science 56: 3–26.

Steel, Daniel 2005b: Mechanism and functional hypotheses in social science. Philosophy of Science 72: 941–952.

Steel, Daniel 2006a: Comment on Hausman & Woodward: On the causal Markov condition. British Journal for the Philosophy of Science 57: 219–231.

Steel, Daniel 2006b: Methodological individualism, explanation, and invariance. Philosophy of the Social Sciences 36: 440–463.

Steel, Daniel 2007: With or without mechanisms. A reply to Weber. Philosophy of the Social Sciences 37: 360–365.

Steel, Mike, Jotun Hein 2006: Reconstructing pedigrees: A combinatorial perspective. Journal of Theoretical Biology 240: 360–367.

Stefanski, Leonard A., Dennis D. Boos 2002: The calculus of M-estimation. American Statistician 56: 29–38.

Stewart, M.B. 1983: On least squares estimation when the dependent variable is grouped. Review of Economic Studies 50: 737–753.

Steyer, Rolf, Sigfried Gabler, Alina A. von Davier, Christof Nachtigall, Thomas Buhl 2000a: Causal regression models I: Individual and average causal effects. Methods of Psychological Research Online 5, No. 2: 39–71.

Steyer, Rolf, Sigfried Gabler, Alina A. von Davier, Christof Nachtigall 2000b: Causal regression models II: Unconfoundedness and causal unbiasedness. Methods of Psychological Research Online 5, No. 3: 55–87.

Steyer, Rolf, Christof Nachtigall, Olivia Wüthrich-Martone, Katrin Kraus 2002: Causal regression models III: Covariates, conditional, and unconditional average causal effects. Methods of Psychological Research Online 7, No. 1: 41–68.

Stocké, Volker, Birgit Becker 2004: Determinanten und Konsequenzen der Umfrageeinstellung. Bewertungsdimensionen unterschiedlicher Umfragesponsoren und die Antwortbereitschaft der Befragten. ZUMA-Nachrichten 54: 89–116.

Stoica, Petre, Luzhu Xu, Jian Li 2005: A new type of parameter estimation algorithm for missing data problems. Statistics & Probability Letters 75: 219–229.

Stolzenberg, Ross M., Daniel A. Relles 1997: Tools for intuition about sample selection bias and its correction. American Sociological Review 62: 494–507.

Strassen, Volker 1964: Meßfehler und Information. Zeitschrift für Wahrscheinlichkeitstheorie 2: 273–305.

Strasser, Helmut 1985: Mathematical Theory of Statistics. Statistical Experiments and Asymptotic Decision Theory. Berlin: Walter de Gruyter.

Strengmann-Kuhn, Wolfgang 1999: Armutsanalysen mit dem Mikrozensus? Pp. 376–402 in: Paul Lüttinger (ed) 1999: Sozialstrukturanalysen mit dem Mikrozensus. ZUMA-Nachrichten Spezial Band 6: Mannheim: ZUMA.

Strickland, Pamela A. Ohman, Benjamin F. Crabtree 2007: Modelling effectiveness of internally heterogeneous organizations in the presence of survey non-response: An application to the ULTRA study. Statistics in Medicine 26: 1702–1711.

Strotz, R.H., Hermann O.A. Wold 1960: Recursive vs. nonrecursive systems: An attempt at synthesis. Econometrica 28: 417–427.

Struthers, Cyntha A., John D. Kalbfleisch 1986: Misspecified proportional hazard models. Biometrika 73: 363–369.

Stubbendick, Amy L., Joseph G. Ibrahim 2003: Maximum likelihood methods for nonignorable missing responses and covariates in random effects models. Biometrics 59: 1140–1150.

Stubbendick, Amy L., Joseph G. Ibrahim 2006: Likelihood-based inference with nonignorable missing responses and covariates in models for discrete longitudinal data. Statistica Sinica 16: 1143–1167.

Sturmfels, Bernd 2002: Solving Systems of Polynomial Equations. CBMS Regional Conferences Series 97. Rhode Island: American Mathematical Society.

Stute, Wolfgang 1995: The central limit theorem under random censorship. Annals of Statistics 23: 422–439.

Stute, Wolfgang, J.-L. Wang 1993: The strong law under random censorship. Annals of Statistics 21: 1591–1607.

Subramanian, Sundarraman 2003: Semiparametric transformation models and the missing information principle. Journal of Statistical Planning and Inference 115: 327–348. Correction: Journal of Statistical Planning and Inference 126 (2004): 397–399.

Subramanian, Sundarraman 2004: The missing censoring-indicator model of random censorship. Pp. 123–141 in: N. Balakrishnan, C.R. Rao (eds) 2004: Handbook of Statistics 23. Advances in Survival Analysis. Amsterdam: Elsevier.

Subramanian, Sundarraman 2006: Survival analysis for the missing censoring indicator model using kernel density estimation techniques. Statistical Methodology 3: 125–136.

Subramanian, Sundarraman 2007: Censored median regression and profile empirical likelihood. Statistical Methodology 4: 493–503.

Sullivant, Seth 2005: Small contingency tables with large gaps. SIAM Journal of Discrete Mathematics 18: 787–793.

Sun, Dongchu, Xiaoqian Sun 2006: Estimation of multivariate normal covariance and precision matrices in a star-shape model with missing data. Journal of Multivariate Analysis 97: 698–719.

Sun, Jianguo 2006: The Statistical Analysis of Interval-censored Failure Time Data. Berlin: Springer.

Sun, Jiayang, Michael Woodroofe 1997: Semi-parametric estimates under biased sampling. Statistica Sinica 7: 545–575.

Sung, Yun Ju, Charles J. Geyer 2007: Monte Carlo likelihood inference for missing data models. Annals of Statistics 35: 990–1011.

Suppes, Patrick 1970: A Probabilistic Theory of Causality. Amsterdam: North–Holland.

Suppes, Patrick 1999: The noninvariance of deterministic causal models. Synthese 121: 181–198.

Suppes, Patrick, Mario Zanotti 1996: Foundations of Probability with Applications. Selected Papers 1974–1995. Cambridge: Cambridge University Press.

Sweeting, Trevor, Samer Kharroubi 2005: Application of a predictive distribution formula to Bayesian computation for incomplete data models. Statistics and Computing 15:167–178.

Tan, Ming, Hong-Bin Fang, Guo-Liang Tian, Peter J. Houghton 2005: Repeated-measures models with constrained parameters for incomplete data in tumor xenograft experiments. Statistics in Medicine 24: 109–119.

Tan, Zhiqiang 2006a: Regression and weighting methods for causal inference using instrumental variables. Journal of the American Statistical Association 101: 1607–1618.

Tan, Zhiqiang 2006b: A distributional approach for causal inference using propensity scores. Journal of the American Statistical Association 101: 1619–1637.

Tang, Gong, Roderick J.A. Little, Trivellore E. Raghunathan 2003: Analysis of multivariate missing data with nonignorable nonresponse. Biometrika 90: 747–764.

Tang, Linqi, Juwon Song, Thomas R. Belin, Jürgen Unützer 2005: A comparison of imputation methods in a longitudinal randomized clinical trial. Statistics in Medicine 24: 2111–2128.

Tanner, Martin A. 1993^2: Tools for Statistical Inference. Methods for the Exploration of Posterior Distributions and Likelihood Functions. Berlin: Springer.

Tarima, Sergey S., Svetla S. Slavova, Travis A. Fritsch, Laura M. Hall 2007: Probability estimation when some observations are grouped. Statistics in Medicine 26: 1745–1761.

Tavaré, Simon 2004: Ancestral inference in population genetics. Pp. 1–188 in: Simon Tavaré, Ofer Zeitouni (eds) 2004: Lectures on Probability Theory and Statistics: Ecole d'Eté de Probabilités de Saint-Flour XXXI - 2001. Berlin: Springer.

Tchernis, Rusty, Marcela Horvitz-Lennon, Sharon-Lise T. Normand 2005: On the use of discrete choice models for causal inference. Statistics in Medicine 24: 2197–2212.

Thiébaut, Rodolphe, Hélene Jacqmin-Gadda, Abdel Babiker, Daniel Commenges, The CASCADE Collaboration 2005: Joint modelling of bivariate longitudinal data with informative dropout and left censoring with applications to the evolution of CD4+ cell count and HIV RNA viral load in response to treatment of HIV infection. Statistics in Medicine 24: 65–82.

Thompson, Elizabeth A., Charles J. Geyer 2007: Fuzzy *p*-values in latent variable problems. Biometrika 94: 49–60.

Thorisson, Hermann 2000: Coupling, Stationarity, and Regeneration. Berlin: Springer.

Tian, Lu, Stephen W. Lagakos 2006: Analysis of a partially observed binary covariate process and a censored failure time in the presence of truncation and competing risks. Biometrics 62: 821–828.

Tien, Hsiao-Chuan, Pranab Kumar Sen 2002: A proportional hazards model for bivariate survival data under interval censoring. Sankh*ā* A 64: 409–428.

Tillé, Yves 2006: Sampling Algorithms. Berlin: Springer.

Tong, Xingwei, Zhongguo Zheng, Zhi Geng 2005: Identifiability, stratification and minimum variance estimation of causal effects. Statistics in Medicine 24: 2937–2952.

Troxel, Andrea B., Stuart R. Lipsitz, David P. Harrington 1998: Marginal models for the analysis of longitudinal measurements with non-ignorable non-monotone missing data. Biometrika 85: 661–672.

Troxel, Andrea B., Guoguang Ma, Daniel F. Heitjan 2004: An index of local sensitivity to nonignorability. Statistica Sinica 14: 1221–1237.

Tryfos, Peter 2004: The Measurement of Economic Relationships. Dordrecht: Kluwer.

Tsai, Wei-Yann, John Crowley 1985: A large sample study of generalized maximum likelihood estimators from incomplete data via self-consistency. Annals of Statistics 13: 1317–1334. Correction: Annals of Statistics 1990. 18: 470.

Tsai, Wei-Yann, John Crowley 1998: A note on nonparametric estimators of the bivariate survival function under univariate censoring. Biometrika 85: 573–580.

Tsiatis, Anastasios A. 2006: Semiparametric Theory and Missing Data. Berlin: Springer.

Tsodikov, A. 2003: Semiparametric models: A generalized self-consistency approach. Journal of the Royal Statistical Society B 65: 759–774.

Tu, X.M., J. Zhang, J. Kowalski, J. Shults, C. Feng, W. Sun, W. Tang 2007: Power analyses for longitudinal study designs with missing data. Statistics in Medicine 26: 2958–2981.

Tuma, Nancy B., Michael T. Hannan 1984: Social Dynamics. Orlando: Academic Press.

Vaihinger, Hans 1911/1923: Die Philosophie des Als Ob. System der theoretischen, praktischen und religiösen Fiktionen der Menschheit auf Grund eines idealistischen Positivismus. Leipzig: Meiner (Volksausgabe 1923. Zuerst: 1911).

van Buuren, Stef 2007: Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research 16: 219–242.

van de Geer, Sara 2000: Empirical Processes in M-Estimation. Cambridge: Cambridge University Press.

van den Berg, Gerard J., Maarten Lindeboom 1998: Attrition in panel survey data and the estimation of multi-state labor market models. The Journal of Human Resources 33: 458–478.

van den Berg, Gerard J., Maarten Lindeboom, Peter J. Dolton 2006: Survey non-response and the duration of unemployment. Journal of the Royal Statistical Society A 169: 585–604.

van der Laan, Mark J. 1995a: Efficient and Inefficient Estimation in Semiparametric Models. Amsterdam: CWI Tracts.

van der Laan, Mark J. 1995b: An identity for nonparametric maximum likelihood estimator in missing data and biased sampling models. Bernoulli 1: 335–341.

van der Laan, Mark J. 1996: Efficient estimation in the bivariate censoring model and repairing NPMLE. Annals of Statistics 24: 596–627.

van der Laan, Mark J. 1997: Nonparametric estimators of the bivariate survival function under random censoring. Statistica Neerlandica 51: 178–200.

van der Laan, Mark J. 2006: Statistical inference for variable importance. The International Journal of Biostatistics 2: Article 2.

van der Laan, Mark J., Richard D. Gill 1999: Efficiency of NPMLE in nonparametric missing data models. Mathematical Methods in Statistics 8: 251–276.

van der Laan, Mark J., Alan Hubbard, Nicholas P. Jewell 2007: Estimation of treatment effects in randomized trials with non-compliance and a dichotomous outcome. Journal of the Royal Statistical Society B 69: 463–482.

van der Laan, Mark J., Alan Hubbard, James M. Robins 2002: Locally efficient estimation of a multivariate survival function in longitudinal studies. Journal of the American Statistical Association 97: 494–597.

van der Laan, Mark J., Maya L. Petersen, Marshall M. Joffe 2005: History adjusted marginal structural models and statically-optimal dynamic treatment regimens. The International Journal of Biostatistics 1: Article 4.

van der Laan, Mark J., Maya L. Petersen 2007: Causal effect models for realistic individualized treatment and intention to treat rules. The International Journal of Biostatistics 3: Article 3.

van der Laan, Mark J., James M. Robins 2003: Unified Methods for Censored Longitudinal Data and Causality. Berlin: Springer.

van der Laan, Mark J., Daniel Rubin 2006: Targeted maximum likelihood learning. The International Journal of Biostatistics 2: Article 11.

van der Linde, Angelika 2004: On the association between a random parameter and an observable. Test 13: 85–111.

van der Vaart, Aad W. 1998: Asymptotic Statistics. Cambridge: Cambridge University Press.

van der Vaart, Aad W. 2004: On Robins' formula. Statistics & Decisions 22: 171–200.

van der Vaart, Aad W., Jon A. Wellner 1996: Weak Convergence and Empirical Processes. With Applications to Statistics. Berlin: Springer.

Vanderweele, Tyler J., James Robins 2008: Empirical and counterfactual conditions for sufficient cause interactions. Biometrika 95: 49–61.

van Es, Bert, Chris A.J. Klaassen, Karin Oudshoorn 2000: Survival analysis under cross-sectional sampling: Length-bias and multiplicative censoring. Journal of Statistical Planning and Inference 91: 295–312.

van Ginkel, Joost R., L. Andries van der Ark, Klaas Sijtsma, Jeroen K. Vermunt 2007: Two-way imputation: A Bayesian method for estimating missing scores and questionnaires, and an accurate approximation. Computational Statistics & Data Analysis 51: 4013–4027.

van Keilegom, Ingrid 2004: A note on the nonparametric estimation of the bivariate distribution under dependent censoring. Nonparametric Statistics 16: 659–670.

van Keilegom, Ingrid, Thomas P. Hettmansperger 2002: Inference on multivariate *M* estimators based on bivariate censored data. Journal of the Americal Statistical Association 97: 328–336.

van Lambalgen, Michiel 2001: Conditional quantification, or poor mans probability. Journal of Logic and Computation 11: 295–335.

Vansteelandt, Stijn, Els Goetghebeur 2004: Using potential outcomes as predictors of treatment activity via strong structural mean models. Statistica Sinica 14: 907–925.

Vansteelandt, Stijn, Els Goetghebeur 2005: Sense and sensitivity when correcting for observed exposures in randomized clinical trials. Statistics in Medicine 24: 191–210.

Vansteelandt, Stijn, Els Goetghebeur, Michael G. Kenward, Geert Molenberghs 2006: Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. Statistica Sinica 16: 953–979.

Vansteelandt, Stijn, Andrea Rotnitzky, James Robins 2007: Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. Biometrika 94: 841–860.

van Steen, Kristel, Geert Molenberghs, Geert Verbeke, Herbert Thijs 2001: A local influence approach to sensitivity analysis of incomplete longitudinal ordinal data. Statistical Modelling 1: 125–142.

van Zwet, Erik W. 2004: Laslett´s line segment problem. Bernoulli 10: 377–396.

Vardi, Yehuda 1982: Nonparametric estimation in the presence of length bias. Annals of Statistic 10: 616–620.

Vardi, Yehuda 1985: Empirical distributions in selection bias models. Annals of Statistics 13: 178–203.

Vella, Francis 1998: Estimating models with sample selection bias: A survey. The Journal of Human Resources 33: 127–169.

Vella, Frank 1988: Generating conditional expectations from models with selectivity bias. Economics Letters 28: 97–103.

Vonesh, Edward F., Tom Greene, Mark D. Schluchter 2006: Shared parameter models for the joint analysis of longitudinal data and event times. Statistics in Medicine 25: 143–163.

von Plato, Jan 1994: Creating Modern Probability. Cambridge: Cambridge University Press.

von Wright, Georg Henrik 1972: Explanation and Understanding. Ithaka: Cornell University Press.

Vytlacil, Edward 2002: Independence, monotonicity, and latent index models: An equivalence result. Econometrica 70: 331–341.

Vytlacil, Edward, Neşe Yildiz 2007: Dummy endogeneous variables in weakly separable models. Econometrica 75: 757–779.

Wahed, Abdus S., Anastasios A. Tsiatis 2006: Semiparametric efficient estimation of survival distributions in two-stage randomisation designs in clinical trials with censored data. Biometrika 93: 163–177.

Wainer, Howard (ed) 1986: Drawing Inferences from Self-Selected Samples. Mahwah: Lawrence Erlbaum.

Wainer, Howard 1989 (1992^2): Eelworms, bullet holes, and Geraldine Ferraro: Some problems with statistical adjustment and some solutions (with discussion). Journal of Educational Statistics 14: 121–140. Reprinted on Pp. 129–20 in: Juliett Popper Shaffer (ed) 1992: The Role of Models in Nonexperimental Social Science: Two Debates. Washington: American Educational Research Association and American Statistical Association.

Walley, Peter 1991: Statistical Reasoning with Imprecise Probabilities. New York: Chapman & Hall.

Walley, Peter 1996: Inference from multinomial data: Learning about a bag of marbles (with discussion). Journal of the Royal Statistical Society B 58: 3–57.

Wallner, Anton 2007: Extreme points of coherent probabilities in finite spaces. International Journal of Approximate Reasoning 44: 339–357.

Wang, Cuiling, Myunghee Cho Paik 2006: Efficiencies of methods dealing with missing covariates in regression analysis. Statistica Sinica 16: 1169–1192.

Wang, C.Y., Hua Yun Chen 2001: Augmented inverse probability weighted estimator for Cox missing covariate regression. Biometrics 57: 414–419.

Wang, C.Y., Yijian Huang, Edward C. Chao, Marjorie K. Jeffcoat 2008: Expected estimating equations for missing data, measurement error, and misclassifcation, with application to longitudinal nonignorable missing data. Biometrics 64: 85–95.

Wang, C.Y., Shen-Ming Lee, Edward C. Chao 2007: Numerical equivalence of imputing scores and weighted estimators in regression analysis with missing covariates. Biostatistics 8: 468–473.

Wang, C.Y., Suojin Wang, Lue-Ping Zhao, Shy-Tyan Ou 1997: Weighted semiparametric estimation in regression analysis with missing covariate data. Journal of the American Statistical Association 92: 512–525.

Wang, C.Y., Sharon X. Xie, Ross L. Prentice 2001: Recalibration based on an approximate relative risk estimator in Cox regression with missing covariates. Statistica Sinica 11: 1081–1104.

Wang, Liansheng, Abba M. Krieger 2006: Causal conclusions are most sensitive to unobserved binary covariates. Statistics in Medicine 25: 2257–2271.

Wang, Mei-Cheng 1999: Gap time bias in incident and prevalent cohorts. Statistica Sinica 9: 999–1010.

Wang, Naisyin, James M. Robins 1998: Large-sample theory for parametric multiple imputation procedures. Biometrika 85: 935-948.

Wang, Qihua 2008: Probability density estimation with data missing at random when covariables are present. Journal of Statistical Planning and Inference 138: 568–587.

Wang, Qihua, J.N.K. Rao 2002: Empirical likelihood-based inference under imputation for missing response data. Annals of Statistics 30: 896–924.

Wang, Qihua, Junshan Shen 2008: Estimation and confidence bands of a conditional survival function with censoring indicators missing at random. Journal of Multivariate Analysis 99: 928–948.

Wang, Qihua, Zhihua Sun 2007: Estimation in partially linear models with missing responses at random. Journal of Multivariate Analysis 98: 1470–1493.

Wang, Sijian, Bin Nan, Ji Zhu, David G. Beer 2008: Doubly penalized Buckley-James method for survival data with high-dimensional covariates. Biometrics 64: 132–140.

Wang, Suojin, C.Y. Wang 2001: A note on kernel assisted estimators in missing covariate regression. Statistics & Probability Letters 55: 439–449.

Wang, Weijing 2003: Estimating the association parameter for copula models under dependent censoring. Journal of the Royal Statistical Society B 65: 257–273.

Wang, Weijing, Martin T. Wells 2000: Model selection and semiparametric inference for bivariate failure-time data. Journal of the American Statistical Association 95: 62–76.

Wang, Yan 2005: A semiparametric regression model with missing covariates in continuous-time capture-recapture studies. Australian New Zealand Journal of Statistics 47: 287–297.

Wang, You-Gan, Vincent Carey 2003: Working correlation structure misspecification, estimation and covariate design: Implications for generalised estimating equations performance. Biometrika 90: 29–41.

Wang, Yue, Oliver Bembom, Mark J. van der Laan 2007: Data-adaptive estimation of the treatment specific mean. Journal of Statistical Planning and Inference 137: 1871–1887.

Wasito, Ito, Boris Mirkin 2006: Nearest neighbours in least-squares data imputation algorithms with different missing patterns. Computational Statistics & Data Analysis 50: 926–949.

Weaver, Mark A., Haibo Zhou 2005: An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. Journal of the American Statistical Association 100: 459–469.

Weber, Erik 2007: Social mechanism, causal inference, and the policy relevance of social science. Philosophy of the Social Sciences 37: 348–359.

Wei, L.J., D.Y. Lin, L. Weissfeld 1989: Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. Journal of the American Statistical Association 84: 1065–1073.

Weichselberger, Kurt 2001: Elementare Grundbegriffe einer allgemeinen Wahrscheinlichkeitsrechnung I. Intervallwahrscheinlichkeit als umfassendes Konzept. Heidelberg: Physica.

Wermuth, Nanny, David R. Cox 2008: Distortion of effects caused by indirect confounding. Biometrika 95: 17–33.

White, Halbert 1982: Maximum likelihood estimation of misspecified models. Econometrica 50: 1–25.

White, Halbert 1983: Maximum likelihood estimation of misspecified models. Correction. Econometrica 51: 513

White, Halbert 2006: Time-series estimation of the effets of natural experiments. Journal of Econometrics 135: 527–566.

White, Ian R., Julian P.T. Higgins, Angela M. Wood 2008: Allowing for uncertainty due to missing data in meta-analysis—Part I: Two-stage methods. Statistics in Medicine 27: 711–727.

White, Ian R., Nicky J. Welton, Angela M. Wood, A.E. Ades, Julian P.T. Higgins 2008: Allowing for uncertainty due to missing data in meta-analysis—Part II: Hierarchical models. Statistics in Medicine 27: 728–745.

Whitrow, G.J. 1963: The Natural Philosophy of Time. New York: Harper.

Whittemore, Alice S., Jerry Halpern 1994: Probability of gene identity by descent: Computation and applications. Biometrics 50: 109–117.

Wichert, Laura, Ralf A. Wilke 2008: Simple non-parametric estimators for unemployment duration analysis. Applied Statistics 57: 117–126.

Wilkins, Kenneth J., Garrett M. Fitzmaurice 2006: A hybrid model for nonignorable dropout in longitudinal binary responses. Biometrics 62: 168–176.

Wilkins, Kenneth J., Garrett M. Fitzmaurice 2007: A marginalized pattern-mixture model for longitudinal binary data when nonresponse depends on unobserved responses. Biostatistics 8: 297–305.

Wilks, S.S. 1932: Moments and distributions of estimates of population parameters from fragmentary samples. Annals of Mathematical Statistics 3: 163–195.

Williams, David 1991: Probability with Martingales. Cambridge: Cambridge University Press.

Williams, J.S., Stephen W. Lagakos 1977: Models for censored survival analysis: Constant-sum and variable-sum models. Biometrika 64: 215–224.

Winship, Christopher, Robert D. Mare 1992: Models for sample selection bias. Annual Review of Sociology 18: 327–350.

Witting, Hermann 1985: Mathematische Statistik I. Stuttgart: Teubner.

Witting, Hermann, Ulrich Müller-Funk 1995: Mathematische Statistik II. Stuttgart: Teubner.

Wong, Linda Yuet-Yee, Qiqing Yu 2007: A bivariate interval censorship model for partnership formation. Journal of Multivariate Analysis 98: 370–383.

Wood, Angela M., Ian R. White, Melvyn Hillsdon, James Carpenter 2005: Comparison of imputation and modelling methods in the analysis of a physical activity trial with missing outcomes. International Journal of Epidemiology 34: 89–99.

Wood, Angela M., Ian R. White, Mathew Hotopf 2006: Using number of failed contact attempts to adjust for non-ignorable non-response. Journal of the Royal Statistical Society A 169: 525–542.

Woodroofe, Michael 1085: Estimating a distribution function with truncated data. Annals of Statistics 13: 163–177.

Woodward, James 2003: Making Things Happen. A Theory of Causal Explanation. Oxford: Oxford University Press.

Woodward, James 2004: Counterfactuals and causal explanation. International Studies in the Philosophy of Science 18: 41–72.

Wooldrige, Jeffrey M. 2001: Asymptotic properties of weighted M-estimators for standard stratified samples. Econometric Theory 17: 451–470.

Wooldrige, Jeffrey M. 2005: Violating ignorability of treatment by contolling for too many factors. Econometric Theory 21: 1026–1028.

Wright, David E., Isabelle Bray 2003: A mixture model for rounded data. The Statistician 52: 3–13.

Wu, C.F. Jeff 1983: On the convergence properties of the EM algorithm. Annals of Statistics 11: 95–103.

Wu, C.-S.P., Y. Zubovic 1995: A large-scale Monte Carlo study of the Buckley-James estimator with censored data. Journal of Statistical Computation and Simulation 51: 97–119.

Wu, Lang 2004: Nonlinear mixed-effect models with nonignorable missing covariates. Canadian Journal of Statistics 32: 27–37.

Wu, Lang 2007: A computationally efficient method for nonlinear mixed effects models with nonignorable missing data in time-varying covariates. Computational Statistics & Data Analysis 51: 2410–2419.

Wu, Lang 2008: An approximate method for nonlinear mixed-effects models with nonignorably missing covariates. Statistics & Probability Letters 78: 384–389.

Wun, Lap-Ming, Trena M. Ezzati-Rice, Nuria Diaz-Tena, Janet Greenblatt 2007: On modelling response propensity for dwelling unit (DU) level nonresponse adjustment in the Medical Expenditure Panel Surveys (MEPS). Statistics in Medicine 26: 1875–1884.

Xie, Jun, Chaofeng Liu 2005: Adjusted Kaplan-Meier estimator and log-rank test with inverse probability of tratment weighting for survival data. Statistics in Medicine 24: 3089–3110.

Yang, S. 1994: A central limit theorem for functionals of the Kaplan–Meier estimator. Statistics & Probability Letters 21: 337–345.

Yang, Xiaowei, Thomas R. Belin, W. John Boscardin 2005: Imputation and variable selection in linear regression models with missing covariates. Biometrics 61: 498–506.

Yang, Xiaowei, Steven Shoptaw, Kun Nie, Juanmei Liu, Thomas R. Belin 2007: Markov transition models for binary repeate measures with ignorable and nonignorable missing values. Statistical Methods in Medical Research 16: 347–364.

Yi, Grace Y., Mary E. Thompson 2005: Marginal and association regression models for longitudinal binary data with drop-outs: A likelihood-based approach. Canadian Journal of Statistics 33: 3–20.

Yi, Grace Y., Richard J. Cook 2002: Marginal methods for incomplete longitudinal data arising in clusters. Journal of the American Statistical Association 97: 1071–1080.

Yi, Guosheng, Jianwen Cai, Jinheum Kim 2003: Quantile inference with multivariate failure time data. Biometrical Journal 45: 602–617.

Yin, Li, Rolf Sundberg, Xiaoqin Wang, Donald Rubin 2006: Control of confounding through secondary samples. Statistics in Medicine 25: 3814–3825.

Yu, Jianqi, K. Krishnamoorthy, Maruthy K. Pannala 2006: Two-sample inference for normal mean vectors based on monotone missing data. Journal of Multivariate Analysis 97: 2162–2176.

Yu, Menggang, Bin Nan 2006: A revisit of semiparametric regression models with missing data. Statistica Sinica 16: 1193–1212.

Yu, Qiqing, George Y.C. Wong 2003: The semi-parametric MLE in linear regression with right censored data. Journal of Statistical Computation & Simulation 73: 833–848.

Yu, Qiqing, George Y.C. Wong, Menggang Yu 2007: Buckley-James-type of estimators under the classical case cohort design. Annals of the Institute of Statistical Mathematics 59: 675–695.

Yu, Zhuo, Mark van der Laan 2006: Double robust estimation in longitudinal marginal structural models. Journal of Statistical Planning and Inference 136: 1061–1089.

Yu, Zhangsheng, Xihong Lin 2008: Nonparametric regression using locl kernel estimating equations for correlated failure time data. Biometrika 95: 123–137.

Yuan, Ying, Roderick J.A. Little 2007a: Model-based estimates of the finite population mean for two-stage cluster samples with unit non-response. Applied Statistics 56: 79–97.

Yuan, Ying, Roderick J.A. Little 2007b: Parametric and semiparametric model-based estimates of the finite population mean for two-stage cluster samples with item nonresponse. Biometrics 63: 1172–1180.

Yun, Sung-Cheol, Youngjo Lee 2006: Robust estimation in mixed linear models with non-monotone missingness. Statistics in Medicine 25: 3877–3892.

Yun, Sung-Cheol, Youngjo Lee, Michael G. Kenward 2007: Using hierarchical likelihood for missing data problems. Biometrika 94: 905–919.

Zaffalon, Marco 2002: Exact credal treatment of missing data. Journal of Statistical Planning and Inference 105: 105–122.

Zaffalon, Marco 2005: Conservative rules for predictive inference with incomplete data. 4th International Symposium on Imprecise Probabilities and Their Application, Pittsburgh.

Zaslavsky, Alan M. 2007: Using hierarchical models to attribute sources of variation in consumer assessments of health care. Statistics in Medicine 26: 1885–1900.

Zeng, Donglin 2004: Estimating marginal survival function by adjusting for depending censoring using many covariates. Annals of Statistics 32: 1533–1555.

Zeng, Donglin, Dan Yu Lin 2007: Maximum likelihood estimation in semiparametric regression models with censored data (with discussion). Journal of the Royal Statistical Society B 69: 507–564.

Zeng, Donglin, Dan Yu Lin 2007a: Efficient estimation for the accelerated failure time model. Journal of the American Statistical Association 102: 1387–1396.

Zeng, Donglin, Dan Yu Lin 2008: Efficient resampling methods for non-smooth estimation functions. Biostatistics 9: 355–363.

Zeng, Donglin, Dan Yu Lin, Xihong Lin 2008: Semiparametric transformation models with random effects for clustered failure time data. Statistica Sinica 18: 355–377.

Zhang, Cun-Hui 2005: Estimation of sums of random variables: Examples and information bounds. Annals of Statistics 33: 2022–2041.

Zhang, Jiajia, Yingwei Peng 2007: An alternative estimation method for the accelerated failure time frailty model. Computational Statistics & Data Analysis 51: 4413–4423.

Zhang, Jiameng, Daniel F. Heitjan 2006: A simple local sensitivity analysis tool for nonignorable coarsening: Application to dependent censoring. Biometrics 62: 1260–1268.

Zhang, Jiameng, Daniel F. Heitjan 2007: Impact of nonignorable coarsening on Bayesian inference. Biostatistics 8: 722–743.

Zhang, Li-Chun 2004: Nonparametric Markov chain bootstrap for multiple imputation. Computational Statistics & Data Analysis 45: 343–353.

Zhang, Li-Chun 2005: On the bias in gross labour flow estimates due to nonresponse and missclassification. Journal of Official Statistics 21: 591–604.

Zhang, Paul 2003: Multiple imputation: Theory and method. International Statistical Review 71: 581–592.

Zhang, Paul 2005: Multiple imputation of missing data with ante-dependence covariance structure. Journal of Applied Statistics 32: 141–155.

Zhang, Zhigang, Liuquan Sun, Jianguo Sun, Dianne M. Finkelstein 2007: Regression analysis of failure time data with informative interval censoring. Statistics in Medicine 26: 2533–2546.

Zhang, Zhiwei, Howard E. Rockette 2005: On maximum likelihood estimation in parametric regression with missing covariates. Journal of Statistical Planning and Inference 134: 206–223.

Zhang, Zhiwei, Howard E. Rockette 2006: Semiparametric maximum likelihood for missing covariates in parametric regression. Annals of the Institute of Statistical Mathematics 58: 687–706.

Zhang, Zhiwei, Howard E. Rockette 2007: An EM algorithm for regression analysis with incomplete covariate information. Journal of Statistical Computation and Simulation 77: 163–173.

Zhao, Yichuan, Feiming Chen 2008: Empirical likelihood inference for censored median regression model via nonparametric kernel estimation. Journal of Multivariate Analysis 99: 215–231.

Zhou, Mai 2005: Empirical likelihood analysis of the rank estimator for the censored accelerated failure time model. Biometrika 92: 492–498.

Zhou, Mai, Gang Li 2008: Empirical likelihood analysis of the Buckley-James estimator. Journal of Multivariate Analysis 99: 649–664.

Zhou, Xiuqing, Jinde Wang 2005: A genetic method of LAD estimation for models of censored data. Computational Statistics & Data Analysis 48: 451–466.

Zhu, Hong-Tu, Sik-Yum Lee, Bo-Cheng Wei, Julie Zhou 2001: Case-deletion measures for models with incomplete data. Biometrika 88: 727–737.

Zhu, Hong-Tu, Sik-Yum Lee 2001: Local influence for incomplete-data models. Journal of the Royal Statistical Society B 63: 111–126.

Ziegler, Günter M. 1995: Lectures on Polytopes. Berlin: Springer.

Zieliński, Ryszard 2004: Effective WLLN, SLLN, and CLT in statistical models. Applicationes Mathematicae 31: 117–125.