Statistical Methods in Sociological Research of Education

G. Rohwer

March 2012

sre.tex March 2012

Contents

Chapter 1

Introduction

The text is partitioned into chapters and sections as indicated in the table of contents. A further subdivision into subsections is provided at the beginning of each chapter.

1.1 Statistical methods

1. Statistical methods vs. case studies
2. Units of analysis
3. Two goals of statistical methods
4. Statistical descriptions

1.2 Educational processes

1. Learning frames
2. Frame-related educational processes
3. Overarching educational processes
4. Construction of statistical variables
5. Transitions and Sequences

1.3 Educational inequalities

1. Single learning frames
2. Inequalities in transitions
3. How to understand opportunities?
4. Measures of educational attainment
5. Inequalities between educational processes

1.4 Empirical illustrations

1. Description of the data
2. Inequalities in schooling outcomes
3. Comparing children with their parents
4. Dependencies on parents' educational level
5. Considerations in terms of stages

This introductory chapter begins with a few general remarks about statistical methods, and distinguishes between two goals of such methods: statistical descriptions of sets of units to which the data relate, and finding rules for prediction and explanation.

In the second section, I consider the question of how to conceptualize educational processes. I propose a general formal framework which is based on a notion of 'learning frames', and briefly show how this framework can be used for defining statistical variables suitable for describing educational processes.

The third section distinguishes four aspects of educational inequality: inequalities (between individuals and between groups) in single learning frames, inequalities in transitions, inequalities in outcomes of educational processes, and inequalities between educational processes. Using data on schooling outcomes from the ALLBUS, some of these aspects of educational inequality are illustrated in the fourth section.

1.1 Statistical methods

1. Statistical methods vs. case studies

A few essential characteristics of statistical methods can best be characterized by contrasting statistical methods with case studies.[1]

- Case studies make in-depth investigations of a few identifiable cases; statistical methods are based on observations of a 'large' number of cases (so that statements about frequencies become sensible).
- In contrast to case studies, statistical methods presuppose the definition of variables (which then determine what counts as a relevant observation).
- Case studies concern identifiable cases. Statistical methods do not require knowledge of identifiable cases; their input eventually only consists of frequency distributions (given the presupposed variables). This entails: Units having identical values in all variables are statistically equal.

Both approaches are complementary. I focus on statistical methods.

2. Units of analysis

Statistical methods deal with 'units of analysis'. I distinguish the following kinds of units: individuals (= individual human beings), situations, events, institutions (including organizations), collectives (sets of units of some kind). It is important that there are different ways of referring to units:

a) identifiable units,

b) quantified units (members of a definite reference set),

c) generic units (not members of a set, defined only by values of variables).

Institutions can be considered in two different ways: as units of analysis, or as conditions for individual processes. The second approach allows one

[1] I here follow Gerring's (2004) understanding of case study research.

- to view institutions in parallel to several other kinds of social environments (for educational processes), and
- include references to properties of institutions into models for individual-based outcome variables.

In both approaches, institutions are conceived of not as identifiable units, but as defined by properties.

3. Two goals of statistical methods

I distinguish two goals of statistical methods:

a) Description of statistical facts. Such facts concern frequency distributions defined for a sample or population (or quantities derived from such distributions).

b) Finding rules for relationships between variables which can be used for explanations and predictions.

Correspondingly, there are two kinds of statistical generalization:

- descriptive generalizations (from sample to population), and
- modal generalizations (from data to rules);

and two kinds of statistical models: descriptive models (for quantified units), and analytical models (for generic units). This will be further discussed in Chapter 2.

4. Statistical descriptions

Statistical descriptions are derived from *statistical variables*. There are two prerequisites:

a set of units (= reference set of a statistical variable), and

a conceptual framework for characterizing the units (= property space of a statistical variable).

Formal representation:

$$X : \Omega \longrightarrow \mathcal{X}$$

X is the name of the variable, Ω is the reference set, and $\mathcal{X}$ is the variable's property space (domain). The notation is intended to show that statistical variables are functions, different from logical variables. As will be argued in **2.1.7**, they are also different from random and modal variables. In particular, one should avoid confusion between statistical variables and their property spaces.

Statistical variables can consist of several components. The distinction between one- and multidimensional statistical variables is not essential, however. It is always assumed that values of statistical variables have a numerical representation (single real numbers, or m-tupels if a variable consists of m components).

The frequency distribution of a statistical variable, say X, will be denoted by $\mathrm{P}[X]$, special values will be denoted by $\mathrm{P}(X = x)$, that is, the proportion of members of Ω with $X = x$. Further notations will be introduced when needed.

1.2 Educational processes

Education is a process that develops in time. In order to apply statistical methods, individual educational processes must be defined in terms of variables. This section introduces a formal framework for defining such variables.

1. Learning frames

As a starting point, I take up the idea that education takes place in a wide variety of 'learning environments' (Bäumer et al. 2011). Formal representations of such environments will be called *learning frames* and symbolically denoted by σ. There definition should include:

a) characterization of the institutional (not necessarily 'formal') context;

b) characterization of the kind of capabilities that can be learned;

c) characterization of the ways by which these capabilities can be acquired;

d) specification of entry requirements (if any);

e) definition of a property space, say $\mathcal{Y}_\sigma$, for the characterization of what has been learned.

The general notion of learning frames covers both broad (e.g. 'elementary school') and fine-grained (e.g. 'learning reading in first grade elementary school') definitions. This must be taken into account when defining the outcome space $\mathcal{Y}_\sigma$. In order to represent outcomes of broadly defined learning frames it could be sensible to use outcome spaces that consist of two or more domain-specific components.

A further question concerns the source of 'observed' outcome values. There are two approaches: One can use the evaluations that are generated in the learning frames (e.g. school certificates), or one can employ external tests developed and applied by a researcher. Choosing one or the other approach depends on the research interest. The first approach corresponds to an interest in learning how educational outcomes are actually generated and evaluated in a society. The second approach corresponds to an interest in developing scientific evaluations.

2. Frame-related educational processes

For each individual participating in a learning frame, say σ, one can think of an educational process that generates a particular outcome, represented by a value in $\mathcal{Y}_\sigma$. This will be called an *individual frame-related educational process* (in contrast to educational processes that extend over several learning frames).

Statistical investigations of such processes start from defining a set of individuals participating in a learning frame of the same type.[2] This allows one to define a statistical variable

$$Y_\sigma : \Omega \longrightarrow \mathcal{Y}_\sigma$$

where Ω denotes a set of individuals participating in the learning frame σ. Of course, it will often be sensible to use a cohort approach where all members of Ω enter the learning frame in the same temporal location (e.g., calendar year).

3. Overarching educational processes

It is more complicated to set up a formal framework for overarching educational processes which extend over several learning frames. My proposal starts from a set of learning frames, say Σ, which are to be considered in an investigation. (As mentioned, broad or fine-grained definitions of learning frames can be used.) An individual educational process, for an individual ω, can then be defined as a sequence of sets

$$\mathcal{S}_\tau(\omega) \subseteq \Sigma$$

containing the learning frames (elements of Σ) in which ω participates at age τ ($\tau = 0, 1, 2, \ldots$). In order to allow easy comparisons of individual educational processes, I take τ to correspond to the calendar year in which ω has the τth birthday ('demographic age').

Since participation in a learning frame can extend over two or more years, it is also useful to define subsets

$$\mathcal{S}^c_\tau(\omega) \subseteq \mathcal{S}_\tau(\omega)$$

consisting of learning frames which ω completes at age τ. There is then,

[2] If not explicitly said otherwise, I use the term 'learning frame' in a generic sense, that is, they are taken as types, not as particular institutions.

for each $\sigma \in \mathcal{S}^c_\tau(\omega)$, a value $Y_\sigma(\omega) \in \mathcal{Y}_\sigma$ which provides information about how ω has completed the learning frame.

4. Construction of statistical variables

The proposed formal framework can be used as a starting point for both statistical descriptions and analytical models. If to be used for statistical descriptions, one needs to specify a set of individuals, say Ω, whose educational processes have been observed and are to be described. So the question arises how to demarcate such sets.

Most often it is sensible to define these sets as cohorts. There are two possibilities:

a) One specifies a calendar year, t^*, and an age, τ^*, and includes all individuals who are of age τ^* in year t^*. (Of course, one can use a range of years instead of just one single year.)

b) One specifies a calendar year, t^*, and a learning frame, σ^*, and includes all individuals who begin participating in σ^* in year t^*.

Following the second approach, basic statistical variables for the reference set Ω can easily be defined. For all $\sigma \in \Sigma$, $t \geq t^*$, and $\omega \in \Omega$:

$S_{t,\sigma}(\omega) = 1$ if ω participates in σ in year t, otherwise zero.

$S^e_{t,\sigma}(\omega) = 1$ if ω begins participation in σ in year t, otherwise zero.

$S^c_{t,\sigma}(\omega) = 1$ if ω completes participation in σ in year t, otherwise zero.

$Y_{t,\sigma}(\omega)$ takes a value in $\mathcal{Y}_\sigma$ (representing the outcome of ω's participation in σ) if $S^c_{t,\sigma}(\omega) = 1$; otherwise there is no valid value.

To these variables which provide the basic information about an individual educational process one can add any number of further variables representing, for example, age, sex, health condition, family background, and characteristics of the home environment and personal networks.

5. Transitions and Sequences

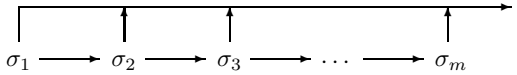
The proposed framework can be used to describe transitions which consist in entering a learning frame. In order to define a transition rate, one needs a risk set and an event set. The event set for entering a learning frame, say σ, can easily be defined:

$$\mathcal{E}_{t,\sigma} := \{\omega \in \Omega \,|\, S^e_{t,\sigma}(\omega) = 1\}$$

The corresponding risk set, $\mathcal{R}_{t,\sigma}$, contains all members of Ω who do not participate in σ in year t and meet the entry conditions for σ (if any). A transition rate can then be defined by $|\mathcal{E}_{t,\sigma}|/|\mathcal{R}_{t,\sigma}|$.

The framework does not presuppose that individuals sequentially participate in learning frames. Any kinds of temporal relationships are possible.

Of course, there often will be entry requirements for the participation in a learning frame which require successful completion of previous learning frames. In the most simple case, there is a linear ordering where successful completion of σ_t is a sufficient precondition for entering σ_{t+1} (Mare 1980, 1981). Graphically:



More complicated sequential schemes have been used, among others, by Hansen (1997), Breen and Jonsson (2000), Goldrick-Rab (2006), Tieben and Wolbers (2010), Schneider and Tieben (2011).

1.3 Educational inequalities

Educational inequalities can be viewed in different ways. In this section I briefly consider four approaches.

1. Single learning frames

I begin with notions of educational inequality which relate to a variable $Y_\sigma : \Omega \longrightarrow \mathcal{Y}$ representing the educational outcome of a single learning frame (σ). A first notion of inequality directly relates to Y_σ and considers the amount of variation in its values as representing educational inequality. Suitable measures depend on $\mathcal{Y}$, the property space of Y_σ. For example:

a) In the most simple case one can use the proportion of individuals who successfully completed the learning frame.

b) If Y_σ can be treated as a metric variable having an approximately symmetric distribution, one can use its variance.

c) If Y_σ is a qualitative (or non-metric quantitative) variable with property space $\mathcal{Y} = \{1, \ldots, m\}$, one can use a *diversity index* defined as

$$1 - \sum_{j=1,m} \mathrm{P}(Y_\sigma = j)^2 \tag{1.1}$$

Using a probabilistic model, this quantity can be interpreted as the probability of meeting two individuals having a different educational outcome.[3]

A different approach is concerned with comparing educational outcomes

[3] For further discussion of this index see Agresti and Agresti (1977).

between groups of individuals. The formal framework then consists of a two-dimensional variable, say

$$(X_\sigma, Y_\sigma) : \Omega \longrightarrow \mathcal{X} \times \mathcal{Y}$$

where $\mathcal{X} = \{1, \ldots, q\}$ distinguishes q groups of individuals.[4] The educational inequality between two groups, say x' and x'', then concerns the difference between their educational outcomes. Formally, it is the difference between two conditional distributions:

$$\mathrm{P}[Y_\sigma \,|\, X_\sigma = x'] \quad \text{and} \quad \mathrm{P}[Y_\sigma \,|\, X_\sigma = x'']$$

Again, there are many different possibilities for quantifying differences between two distributions. If Y_σ is a discrete variable, one can use the *dissimilarity index*

$$\frac{1}{2} \sum_{j=1,m} \big| \mathrm{P}(Y_\sigma = j \,|\, X_\sigma = x') - \mathrm{P}(Y_\sigma = j \,|\, X_\sigma = x') \big| \tag{1.2}$$

It can be interpreted as the proportion of individuals in one of the two groups who must attain a different educational outcome in order make the two outcome distributions identical.[5]

2. Inequalities in transitions

I now consider a different kind of inequality that concerns differences, between groups, of entries into learning frames. A simple framework suffices for an explanation of the basic idea. There are three variables:

$$(X, C, Y) : \Omega \longrightarrow \mathcal{X} \times \{0, 1\} \times \{0, 1\}$$

X with property space $\mathcal{X} = \{1, \ldots, q\}$ provides labels of q groups; C and Y are binary variables (σ denotes a learning frame):

$C(\omega) = 1$ if ω has the opportunity to enter σ
$Y(\omega) = 1$ if ω actually enters σ

This framework allows one to make three kinds of comparisons between groups, say x' and x''. A first possibility is to compare

$$\mathrm{P}(C = 1 \,|\, X = x') \quad \text{and} \quad \mathrm{P}(C = 1 \,|\, X = x'')$$

[4] In sociological research of education, the demarcation of groups is often based on student's sex and migration background, and on their parents' educational level and socio-economic status.

[5] This index is a special case of a general class of substitution metrics that rely on evaluations of redistributions for quantifying differences between statistical distributions.

This comparison shows inequalities in having the opportunity for entering the learning frame. A second possibility is to compare

$$\mathrm{P}(Y = 1 \,|\, C = 1, X = x') \quad \text{and} \quad \mathrm{P}(Y = 1 \,|\, C = 1, X = x'')$$

This comparison shows inequalities in actually entering the learning frame, given that there is an opportunity for doing so. Finally, both can be combined using the equality

$$\mathrm{P}(Y = 1 \,|\, X = x) = \mathrm{P}(Y = 1 \,|\, C = 1, X = x)\, \mathrm{P}(C = 1 \,|\, X = x)$$

This decomposition shows how the 'unconditional' inequality results from inequalities in opportunities and inequalities in using an opportunity.

3. How to understand opportunities?

While the formal framework for a consideration of inequalities in transitions is quite simple, there is a real difficulty that concerns the understanding of 'opportunities'. The difficulty is entailed by the following semantical requirement:

> To say that an individual *has* a particular opportunity presupposes that the individual *is aware of* that opportunity and believes in the possibility of making use of the opportunity.

Given this understanding, 'opportunity' is an essentially subjective notion, and one often will not be able to find statistical data representing opportunities in this sense. Instead, one has to rely on 'objective' definitions based, e.g., on formally defined rights. As an example, one can think of formally defined prerequisites for the right of attending a specified school type.

In any case, I propose that one explicitly distinguishes between opportunities (which presuppose that there is a choice) and final outcomes resulting from different opportunities and actual choices. Of course, a corresponding distinction should be made in the talk of 'chances'.

4. Measures of educational attainment

The notions of inequality considered so far concern temporally limited situations: single learning frames or single transitions. Further questions concern inequalities in overarching educational processes. An easy approach refers to final outcomes of such processes, often called 'educational attainment'. There are then three possibilities to construct variables.[6]

[6] I presuppose the understanding of qualitative, quantitative and metric variables proposed in Rohwer and Pötter (2002).

a) One can think in terms of qualitatively different outcomes. Formally, one constructs a set of possible final outcomes, say $\mathcal{Y}$, and treats each element of $\mathcal{Y}$ as a qualitatively different type. $\mathcal{Y}$ can then be used as the property space of a multinomial variable.

b) Starting from $\mathcal{Y}$, one can propose a quantification, that is, a linear ordering of the elements of $\mathcal{Y}$. $\mathcal{Y}$ can then be used as the property space of a quantitative (ordinal) variable.

c) Starting from $\mathcal{Y}$, one can propose a metric, that is, a function which assigns to each pair of elements a distance value. $\mathcal{Y}$ can then be used as the property space of a metric variable. Often, but not necessarily, the introduction of a metric follows a previous quantification. A simple example is using 'years of schooling (and training)' for quantification and differences of such values for the metric.

In any case, having defined a property space for 'educational attainment', one can use the methods discussed in **1.3.1** to describe educational inequalities.

5. Inequalities between educational processes

Finally, one can attempt to directly assess differences between individual educational processes. This is difficult when starting from the very general notion of such processes that was proposed in Section 1.2. There are basically two methods.

a) One can calculate the frequency of each type of observed process. In particular with fine-grained process frames, this is often not informative because of the large number of different process types.

b) One can define a distance function for educational processes. Then, with n different process types, one can calculate the $n \times n$ distance matrix and finally use a method (e.g. MDS or cluster analysis) for presenting its structure.

To achieve informative results, it could be sensible to partition the given set of processes into a relatively small number of 'typical' trajectories. Formally, this means to construct a property space for educational processes, say $\mathcal{Y}_p$. This allows one to define a statistical variable, say Y_p, assigning each individual to a particular educational process type.

To enhance information about educational inequality, one can investigate relationships between Y_p and other variables characterizing the individuals. For example, Goldrick-Rab (2006) used types of college pathways as the dependent variable in a multinomial logit model.

However, while useful in descriptions of educational inequality, it is questionable whether variables representing process types should be used as dependent variables in explanatory models. If one conceives of educational processes as resulting from a multitude of temporally local events, also their explanation should start from these events.

1.4 Empirical illustrations

To illustrate some possibilities to describe aspects of educational inequality, I use data on schooling outcomes from the ALLBUS. This is a general purpose social survey conducted in Germany biannually since 1980. I use a cumulated version of the data set that covers the years from 1980 to 2008.

1. Description of the data

The variable of interest, outcome of schooling, is coded as follows:

1 = without Hauptschulabschluss
2 = Hauptschulabschluss
3 = Realschulabschluss
4 = Fachhochschulreife
5 = Abitur

The same information is available for the father and, since 1984, the mother of the interviewed person. For constructing a single variable for 'parents' outcome of schooling', I use the highest of the two values.

In order to get comparable data for a long range of birth cohorts, I only include individuals who lived in West Germany when interviewed. This also allows going without sampling weights (for different inclusion probabilities in West and East Germany). I finally get data for age cohorts 1908 – 1986. The number of persons with a valid information about schooling outcome is 38461; for 36596 of these persons one also knows the parents' outcome of schooling.

2. Inequalities in schooling outcomes

Figure 1.1 uses the diversity index, defined in (1.1), to describes inequalities in schooling outcomes. The increase in inequality is mainly due to an increasing proportion of persons who attain higher schooling outcomes, starting from a historical situation where most people finished with Hauptschule.

In order to show that there is a tendency towards higher schooling outcomes, I use the following index:

$$d_t := \sum_{j=1,5} (5-j)\,\mathrm{P}(Y_t = j) \tag{1.3}$$

where Y_t represents the schooling outcomes of persons belonging to birth cohort t. This index compares the actual outcome distribution with an

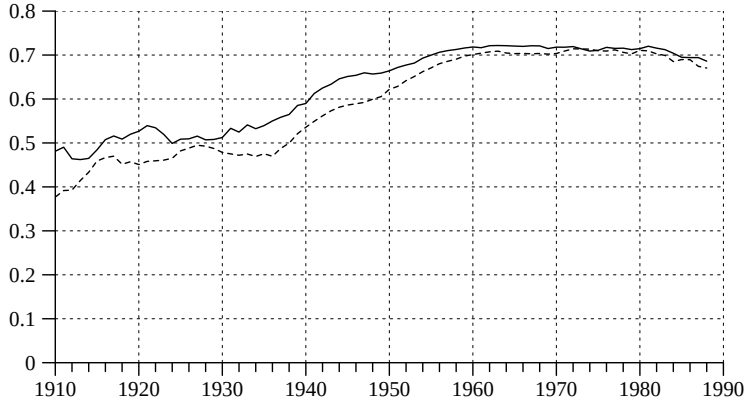


Figure 1.1 Development of the diversity index for schooling outcomes of men (solid line) and women (dashed line) of the birth cohorts shown on the abscissa. Based on data from the Cumulated ALLBUS 1980–2008; smoothed with running means of length 7.

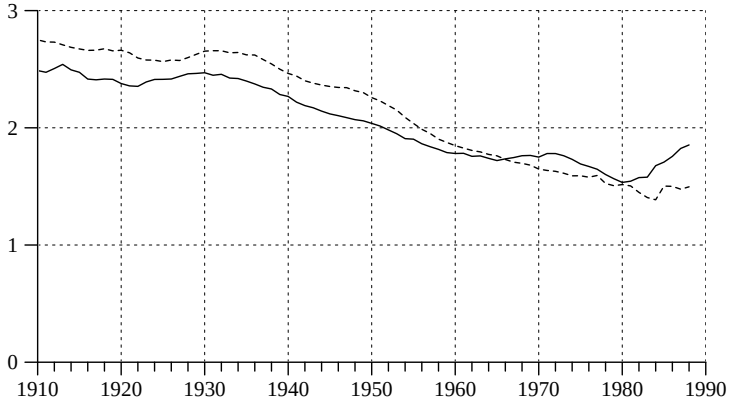


Figure 1.2 Development of the index d_t for schooling outcomes of men (solid line) and women (dashed line) of the birth cohorts shown on the abscissa. Based on data from the Cumulated ALLBUS 1980–2008; smoothed with running means of length 7.

ideal distribution which corresponds to a situation where each person leaves school with an Abitur. The lower the value of d_t, the smaller is the distance between the actual and the ideal distribution of schooling outcomes. The development of this index is shown in Figure 1.2.

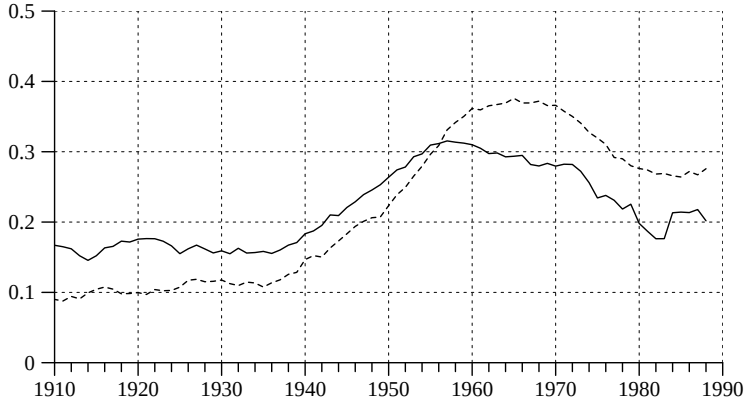


Figure 1.3 Development of the index DI_t for men (solid line) and women (dashed line) of the birth cohorts shown on the abscissa. Based on data from the Cumulated ALLBUS 1980–2008; smoothed with running means of length 9.

3. Comparing children with their parents

I now compare the schooling outcomes of the interviewed persons (in the following also called 'children') with the schooling outcomes of their parents. Starting point is the twodimensional variable (X_t, Y_t) with Y_t recording the schooling outcome of persons belonging to birth cohort t, and X_t recording the schooling outcome of *their* parents.

As a first approach, one can simply compare the distributions of schooling outcomes. That is, one compares $\mathrm{P}[Y_t]$ with $\mathrm{P}[X_t]$. This can be done with a dissimilarity index

$$\mathrm{DI}_t := \frac{1}{2} \sum_{j=1,5} \big|\, \mathrm{P}(Y_t = j) - \mathrm{P}(X_t = j) \,\big| \tag{1.4}$$

The development of this index is shown in Figure 1.3. Large values occur in a historical situation where there is a large proportion of children who attain a higher schooling outcome than their parents. Eventually, when these children become parents of their own children, the index decreases.

4. Dependencies on parents' educational level

A further question concerns how the schooling outcome of children depends on the educational level of their parents. I compare two groups: children whose parents have a Hautschulabschluss and children whose parents have an Abitur. Figure 1.4 shows for these groups the proportion of children

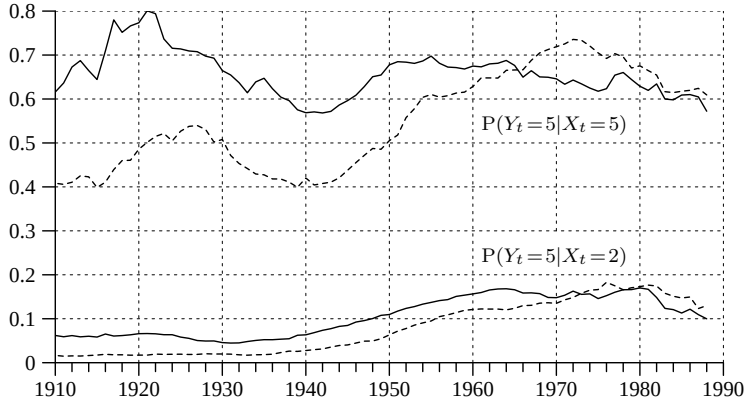


Figure 1.4 Proportion of men (solid lines) and women (dashed lines) who attained Abitur, differentiated w.r.t. the educational level of their parents and w.r.t. birth cohorts shown on the abscissa. Based on data from the Cumulated ALLBUS 1980–2008; smoothed with running means of length 11.

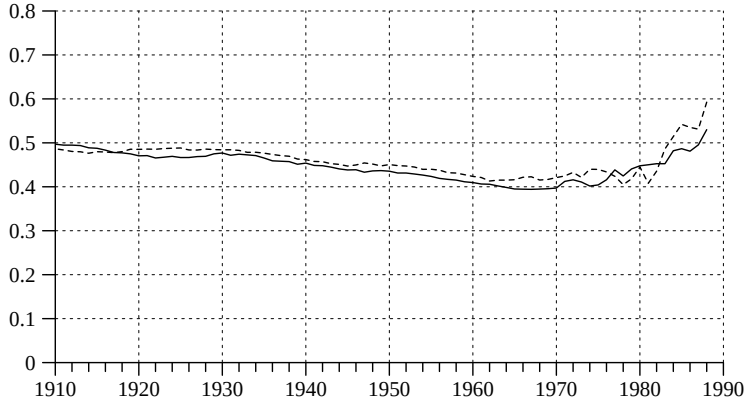


Figure 1.5 Development of the index DI_t^* for men (solid line) and women (dashed line) of the birth cohorts shown on the abscissa. Based on data from the Cumulated ALLBUS 1980–2008; smoothed with running means of length 7.

who attain an Abitur. Formally, the comparison is between

$$\mathrm{P}(Y_t=5\,|\,X_t=2) \quad \text{and} \quad \mathrm{P}(Y_t=5\,|\,X_t=5)$$

Figure 1.4 shows how these proportions have changed across cohorts.

Another possibility is to compare the distribution of schooling outcomes

in the two groups. I use again a dissimilarity index, now defined as

$$\mathrm{DI}_t^* := \frac{1}{2}\sum_{j=1,5}\big|\,\mathrm{P}(Y_t=j|X_t=2)-\mathrm{P}(Y_t=j|X_t=5)\,\big| \tag{1.5}$$

The development of this index is shown in Figure 1.5.

5. Considerations in terms of stages

Inequalities in educational attainment are often discussed in terms of stages (Mare 1980, 1981; Shavit and Blossfeld 1993). One can then consider transition rates between stages. Depending on the definition of educational stages, such rates can relate to transitions which can be actually performed by students, or they are analytical tools for describing inequalities in educational attainment. Only knowing the final schooling outcomes, one cannot reconstruct the actual transitions. The consideration of schooling outcomes in terms of conditional frequencies is nevertheless informative. I use the following stage variables:

$$T_{1t} := \begin{cases} 1 & \text{if } Y_t \geq 3 \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad T_{2t} := \begin{cases} 1 & \text{if } Y_t \geq 4 \\ 0 & \text{otherwise} \end{cases}$$

One can then define the following quantities, all conditional on a specified educational level of the parents:

$\mathrm{P}(T_{1t}=1|X_t=x)$, that is, the proportion of children who attain at least Realschulabschluss in the set of all children.

$\mathrm{P}(T_{2t}=1|X_t=x)$, that is, the proportion of children who attain at least Fachhochschulreife or Abitur in the set of all children.

$\mathrm{P}(T_{2t}=1|T_{1t}=1,X_t=x)$, that is, the proportion of children who attain at least Fachhochschulreife or Abitur in the set of children who attain at least Realschulabschluss.

The following relationship holds between these quantities:

$$\mathrm{P}(T_{2t}=1|X_t=x)=\mathrm{P}(T_{2t}=1|T_{1t}=1,X_t=x)\,\mathrm{P}(T_{1t}=1|X_t=x)$$

The development of these quantities across cohorts is shown for men in Figure 1.6 and for women in Figure 1.7. The comparison is between persons whose parents have Hauptschulabschluss ($X_t=2$, solid lines) and persons whose parents have Abitur ($X_t=5$, dashed lines). Corresponding effects can be defined as differences between the defined quantities, e.g.

$$\mathrm{P}(T_{2t}=1|X_t=5)-\mathrm{P}(T_{2t}=1|X_t=2)$$

represents the total effect of this contrast of parents' educational level on attaining Fachhochschulreife or Abitur.

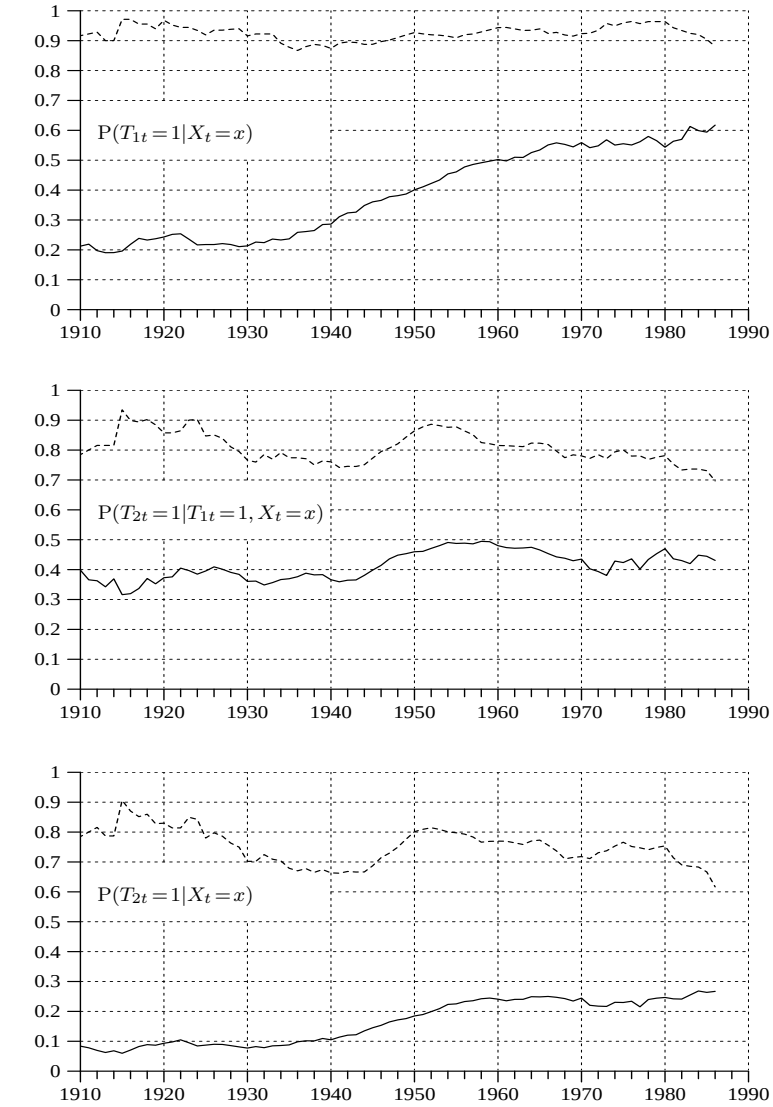


Figure 1.6 Proportion of men who attained at least Realschulabschluss ($T_{1t} = 1$) or Fachhochschulreife/Abitur ($T_{2t} = 1$), dependent on the educational level of their parents (solid line: Hauptschulabschluss, dashed line: Abitur), for birth cohorts shown on the abscissa. Based on data from the Cumulated ALLBUS 1980–2008; smoothed with running means of length 7.

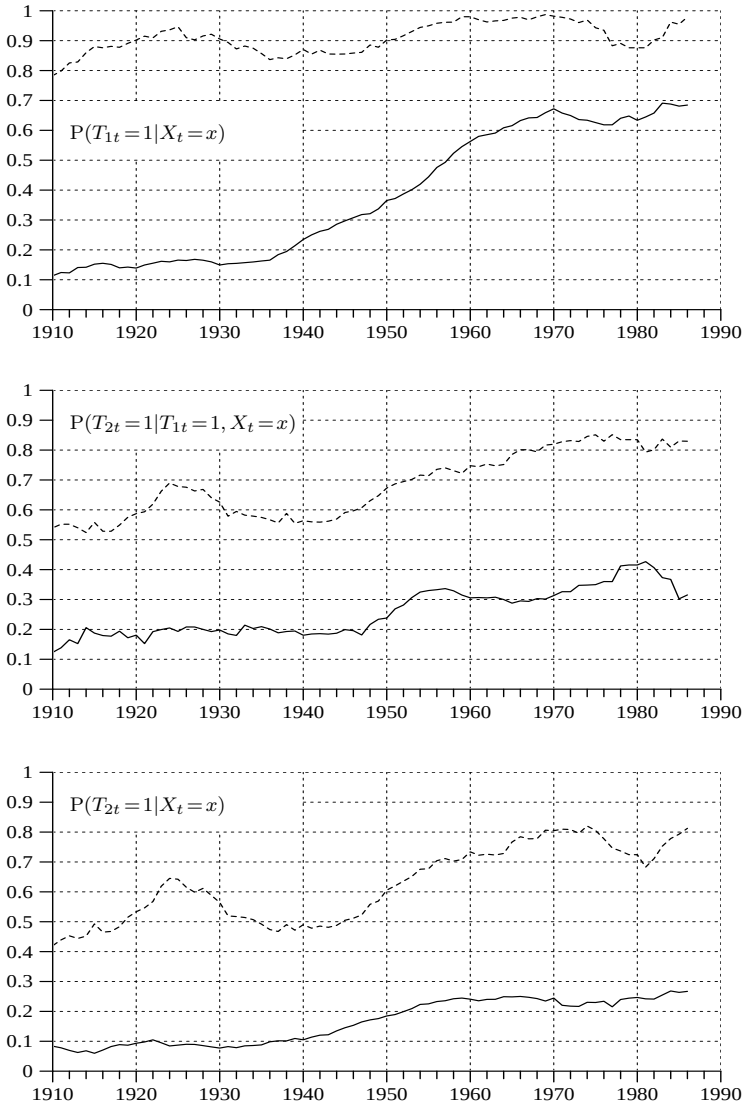


Figure 1.7 Proportion of women who attained at least Realschulabschluss ($T_{1t} = 1$) or Fachhochschulreife/Abitur ($T_{2t} = 1$), dependent on the educational level of their parents (solid line: Hauptschulabschluss, dashed line: Abitur), for birth cohorts shown on the abscissa. Based on data from the Cumulated ALLBUS 1980–2008; smoothed with running means of length 7.

Chapter 2

Descriptive and analytical models

In the introduction, I distinguished two goals of statistical methods: Description of statistical distributions in a specified set of units (a sample or population), and finding rules, most often about relationships between variables, which can be used for predictions and explanations. There correspondingly are two ways of using data:

– descriptive generalization: from descriptive statements about a sample to descriptive statements about a corresponding population, and

– modal generalization: from statistical data to the formulation of more general relationships between variables.

This distinction will be discussed in the first section. I then go on to distinguish two kinds of statistical models. In the second section I very briefly consider descriptive models which aim to describe distributions of statistical variables defined for a sample or population. In the third section I consider analytical models serving to formulate theoretical hypotheses about relationships between variables.

2.1 Two kinds of generalization

1. Descriptive statistical statements

I use the following definition:

> Descriptive statistical statements are statements about the frequency distribution of properties (or quantities derived thereof) in a specified set of units.

As already indicated in the introduction, I use statistical variables as a formal framework. The symbolic notation is

$$X : \Omega \longrightarrow \mathcal{X}$$

X is the name of the variable, Ω is the reference set, a finite set of actually observed or assumed cases, and $\mathcal{X}$ is the property space (domain).

If the reference set consists of not actually observed cases, it is nevertheless required, for descriptive statements, that one can reasonably assume that the cases do exist, or have existed in the past. For example, 'all children who attended Kindergarten in Germany in 2010', but not: 'all children who (possibly) attend Kindergarten in Germany in 2015'.

2. Defining descriptive generalization

Statistical variables provide a useful framework for the definition of descriptive generalization. Starting point is a statistical variable, $X : \Omega \longrightarrow \mathcal{X}$, representing the observations. Ω, the set of observed cases, is then considered as subset of another set, Ω^*, for which one can assume an analogously defined statistical variable:

$$X^* : \Omega^* \longrightarrow \mathcal{X}$$

having the same property space as X.

This framework allows one to define: A *descriptive generalization* consists in using the observed values of X for making descriptive statements about the distribution of X^* in Ω^*. It is noteworthy that the desired generalization has the same linguistic form as the statistical statements derived from the observations; there only is a change in the reference set.

I will not discuss here problems of statistical inference. It is obvious, however, that the justification of a descriptive generalization must be based on the data generating process that has generated the observations. Note that I use the following distinction:

– The term 'data generating process' is used to refer to a process that generates data, that is, information about already existing facts.

– In contrast, when referring to processes that generate new facts (outcomes), I use the term 'fact-generating process'.

As an example think of a learning frame in which students can acquire capabilities of a specified kind, and assume that individual learning results can be captured by values of a variable, say Y. One can firstly think of a fact-generating process in which each student eventually acquires a particular capability. Afterwards, a data-generating process can take place, that is, a process in which a researcher represents students' capabilities by particular values of Y.

3. Limitations of descriptive generalizations

Descriptive generalization intends to enlarge the knowledge about statistical facts, meaning here statistical distributions as they are actually realized in specified populations. This very interest requires a narrow understanding of 'population'.

Limitations become obvious when the justification of descriptive generalization is based on probability sampling. This requires that Ω can be viewed as a probability sample from Ω^*. Consequently, Ω^* can only consist of units having a positive selection probability *when and where* the sample is drawn.

Particular difficulties arise when the interest concerns historical processes. The basic question then is, How to define a population of processes? From a methodological point of view, such populations are best defined as cohorts. However, being interested in descriptive generalizations, this requires to adopt a historical perspective that is confined to mostly completed processes.

A special problem occurs in the NEPS which is based on different samples from different populations: Would it be possible to combine the data in order to get a picture of overarching educational processes? This will not be possible in the form of just one descriptive generalization (since no combined sampling design is available). It might be possible, however, to use the separate samples for modal generalizations (predictive rules) which, taken together, would allow making comprehensive statements about educational processes.

4. Modal generalizations with rules

I now consider a different kind of generalization where the goal is, not a descriptive statement about a set of units, but a predictive rule. I use the term 'rule' in a general sense for statements having the form

If ..., then ...

Different kinds of rules can be distinguished w.r.t. the modalities used in formulating the *then*-part; for example: If ..., then ... is possible, or probable, or necessary, or normatively required.

Empirical research is primarily interested in *predictive rules*. Example: Let ω denote an individual who has finished school in Germany: If at least one of ω's parents has finished school with an Abitur, then it is highly probable that also ω has an Abitur. Note that this is a *generic* rule, meaning that its object is specified only by values of variables.

An important distinction can be made between static and dynamic predictive rules.

- A *static predictive rules* formulates a relationship between properties of a unit. The general form is: If ω has property x, then ω (probably) has property y. This kind of predictive rule is exemplified by the above example.
- A *dynamic predictive rules* relates to a fact-generating process that generates an outcome that is to be predicted.
  Example: If ω (a generically specified individual) regularly participates in the instructions, she will (probably) be successful in the final exam.

5. Formulating predictive rules with variables

In the following, I only consider predictive rules which include a probabilistic qualification of the prediction. When formulating such rules with variables, a first question concerns how to understand the probabilistic qualification. There are two forms:

a) Qualitative: If $X{=}x$, then $Y{=}y$ is probable (in some qualified sense).

b) Quantitative: If $X{=}x$, then $\Pr(Y{=}y) = \ldots$ [a specific, actually given or assumed, numerical value].

Empirical research with statistical methods regularly uses quantitative formulations. (Since there is a formal equivalence of frequency and probability functions, researchers often ignore the conceptual distinction and present their observed frequencies in terms of probabilities.)

The presupposition of quantifiable probabilities allows one to use mathematical functions for formulating the relationship between the *if*- and the *then*-part of the rule. As a general form one can use

$$x \longrightarrow \Pr[Y \mid X{=}x]$$

to be read as a function that assigns to each value x in the domain of X a conditional probability distribution of Y. If Y is a discrete variable, one can also use specific functions having the form

$$x \longrightarrow \Pr(Y{=}y \mid X{=}x)$$

for each value y in the domain of Y. Another often used special form is

$$x \longrightarrow \mathrm{E}(Y \mid X{=}x)$$

which formulates the relationship with conditional expectations of Y.

Starting from such general formulations, one can think of more specific parametric forms. However, whatever the finally chosen functional form, these forms must be distinguished from numerically specified functions which actually allow one to calculate values of the function.

6. Predictive rules vs. descriptive statements

Predictive rules must be distinguished from descriptive statistical statements.

- While descriptive statistical statements concern a reference set of particular units, a predictive rule concerns a generic unit which is only specified by values of variables.
- Correspondingly, there is a conceptual difference between frequencies,

  $$\mathrm{P}(Y{=}y \,|\, X{=}x)$$

  which presuppose a finite reference set, and probabilities,

  $$\Pr(Y{=}y \,|\, X{=}x)$$

  which concern a generic unit. I therefore use different symbols: P for frequencies, and Pr for probabilities.

A random generator can serve to illustrate the distinction. I use 'throwing a die' as an example. The random generator can be defined by a rule, e.g.,

> If the die is thrown, there are six possible outcomes, each can occur with the same probability (1/6).

This rule is to be distinguished from a descriptive statement about frequencies of outcomes in an actually realized set of throws.

Assume the die is thrown 100 times. Results can be represented by a statistical variable $Z : \Omega \longrightarrow \mathcal{Z} := \{1, \ldots, 6\}$. P[$Z$], the distribution of Z, must be distinguished (numerically and conceptually) from the probability distribution which is used in the formulation of the rule describing the random generator.

Since predictive rules are different from descriptive statements (and different from analytical truths), they cannot be true or false. They can only be pragmatically justified, that is, with arguments showing that, and how, a rule can help people in their activities.

7. Statistical and modal variables

The conceptual distinction between descriptive statements and predictive rules suggests to make a corresponding distinction between the kinds of variables involved. As already explained, descriptive statistical statements

are derived from statistical variables which are known, or assumed, to represent realized properties of existing units. This is also true for conditional frequencies: $\mathrm{P}(Y{=}y \,|\, X{=}x)$ is derived from a statistical variable, (X, Y), which is defined for a particular reference set.

When considering instead conditional probabilities, $\Pr(Y{=}y \,|\, X{=}x)$, one must recognize that there are two different conceptual frameworks:

- One can assume that the conditional probabilities are derived from a random variable (X, Y). This understanding presupposes the existence of joint and marginal probability distributions of the two variables.
- The situation is different when conditional probabilities serve to formulate probabilistic predictive rules. In this framework, X is used to formulate a hypothetical assumption, and so it is neither a random variable (having a probability distribution) nor a statistical variable (having a statistical distribution). Consequently, also Y has no unconditional distribution, but can only be viewed as a random variable for specified values of X.

In order to remind of the second context, I speak of *modal variables* and use a special notation: $\ddot{X}$ instead of X, and $\dot{Y}$ instead of Y. The symbolic notation for a probabilistic predictive rule then becomes

$$x \longrightarrow \Pr[\dot{Y} \,|\, \ddot{X}{=}x]$$

8. Applications of predictive rules

Why do we need predictive rules? Part of the answer is obvious: they are required for making observations relevant for predicting possible (future) outcomes. However, can predictive rules also be used for explanations?

An influential tradition has sought to use probabilistic rules for explanations of individual outcomes (often called 'inductive-statistical explanations'). This proposal was followed by a longstanding critical discussion. One of the main points of criticism is easily understandable: That A makes B to some degree probable does not show why B occurred.

The discussion mainly concerns differences between 'explanation' and 'prediction' and can therefore be ignored if one is only interested in predictions. On the other hand, being interested in explanations, one should begin with rethinking the questions that should be answered by an explanation and, in particular, distinguish 'why' and 'how' questions (see, e.g., Cross (1991) and Faye (1999)).

I will not take up this discussion which primarily concerns the explanation of particular outcomes in individual cases. Instead, I briefly consider statistical explanations which are concerned with statistically defined explananda (= statistical distributions or quantities derived from such distributions).

9. Statistical explanations

Statistical explanations, as I use this term here, are concerned with the explanation of statistical distributions. Let $\mathrm{P}[Y]$, the distribution of a statistical variable $Y : \Omega \longrightarrow \mathcal{Y}$, denote the explanandum. In my understanding, a statistical explanation uses two premisses:

- a statistical distribution, $\mathrm{P}[X]$, X being defined for the same reference set Ω, and
- a probabilistic rule: $x \longrightarrow \Pr[\dot{Y} | \ddot{X} = x]$, where $\ddot{X}$ and $\dot{Y}$ correspond, respectively, to X and Y.

The formal part of the explanation then consists in using

$$\Pr(\dot{Y}{=}y) := \sum\nolimits_x \Pr(\dot{Y}{=}y \,|\, \ddot{X}{=}x)\, \mathrm{P}(X{=}x) \tag{2.1}$$

to derive a probability distribution $\Pr[\dot{Y}]$. (This formulation presupposes that all variables are discrete.)

The predictive claim is that $\Pr[\dot{Y}]$ is approximately equal to $\mathrm{P}[Y]$. This claim is trivially valid if the predictive rule is derived from the joint distribution of X and Y. The idea that the predictive rule is a generalization must therefore be taken seriously, and should be explicitly considered.

However, one also must reflect explanatory claims which are entailed neither by the formal framework nor by any particular degree of predictive success. I propose that the following considerations are important.

a) Whether values of explanatory variables can be understood as conditions for processes generating values of the explanandum variable.

b) Whether there are relationships between explanatory variables, and one can distinguish mediating and exogenous explanatory variables.

c) Whether there are potentially important explanatory variables not explicitly considered, and what might follow from their omission.

d) Whether, and to which degree, the rule used in the explanation depends on the particular distribution of the explanatory variables ('distribution-dependent causation').

e) Whether, and to which degree, the rule used in the explanation is historically stable.

2.2 Descriptive models

1. Descriptive models based on statistical Variables

Descriptive models, as understood in this text, are tools for describing distributions of statistical variables. Starting point is a statistical variable, $X : \Omega \longrightarrow \mathcal{X}$, often consisting of several components. A descriptive model aims to describe the distribution of X, denoted by $\mathrm{P}[X]$, or aspects of this distribution, by using a simpler mathematical form.

As an example, one can think of describing the distribution of students' 'ability scores' by a normal distribution (Jackson et al. 2007).

2. Regression models with statistical variables

If X consists of two or more components, one is often interested in descriptions of conditional distributions. This is done with regression functions and regression models. The starting point is given by a two-dimensional statistical variable, say

$$(X, Y) : \Omega \longrightarrow \mathcal{X} \times \mathcal{Y}$$

A general *regression function* is a function

$$x \longrightarrow \mathrm{P}[Y \,|\, X{=}x]$$

which assigns to each value $x \in \mathcal{X}$ the conditional frequency distribution of Y, as given by the statistical variable (data). In order to create a descriptive regression model, one uses a simpler mathematical representation of the conditional distribution, say

$$g(x; \theta) \approx \mathrm{P}[Y \,|\, X{=}x]$$

where θ is a parameter vector. A general *regression model* is then given by the function $x \longrightarrow g(x; \theta)$.

Special regression models are used to represent aspects of $\mathrm{P}[Y \,|\, X{=}x]$. Of widespread use is regression with mean values: $m(x; \theta) \approx \mathrm{M}(Y \,|\, X{=}x)$. As an example, one can think of a linear model,

$$\mathrm{M}(Y \,|\, X{=}x) \approx \alpha + x\beta$$

This model approximates the conditional mean value of Y by a linear function of the values of X that are used as conditions.

3. Descriptive models and descriptive generalizations

Descriptive models are primarily tools for comprehending aspects of complex data sets. As suggested by a famous statistician, R. A. Fisher (1922: 311), this is a primary task of statistical methods:

> Briefly, and in its most concrete form, the object of statistical methods is the reduction of data. A quantity of data, which usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data.

Descriptive models are also useful tools for descriptive generalizations. Since these models relate to statistical variables, there is no conceptual difference whether the reference set is a sample or a population. Starting from a model intended to describe a population allows one to think in terms of estimating its parameters with the information from a sample.

Notice that the term 'estimation' has a clear meaning in this context: It means that one aims to find values of model parameters which are defined by their hypothetical calculation for the complete population. This entails that already their definition depends not only on the specified model, but also on a particular method to calculate its parameters.

2.3 Analytical models

1. Relationships between variables

Analytical models, as understood in this text, are tools for thinking about relationships between variables. The basic formal tool are functions (mathematically understood) which connect variables. So one can speak of 'functional relationships' between variables, and the models are also called 'functional models' (Rohwer 2010).

Two kinds of such functional relationships must be distinguished. Consider two variables, X with domain $\mathcal{X}$ and Y with domain $\mathcal{Y}$.

– A *deterministic* functional relationship consists of a function

$$x \longrightarrow y = f(x)$$

which assigns to each value $x \in \mathcal{X}$ exactly one value $f(x) \in \mathcal{Y}$.

– A *probabilistic* functional relationship consists of a function

$$x \longrightarrow \Pr[Y|X{=}x]$$

which assigns to each value $x \in \mathcal{X}$ a conditional probability distribution.

Notice that $\Pr[Y|X{=}x]$ is itself a function. If Y is a discrete variable, this function can be written as

$$y \longrightarrow \Pr(Y{=}y \,|\, X{=}x) \tag{2.2}$$

to be interpreted as the probability of $Y{=}y$ given that $X{=}x$.

Since (2.2) is formally identical with a probabilistic predictive rule as introduced in **2.1.6**, functional models can also be understood as tools for formulating probabilistic predictive rules.

2. A general notion of functional models

A general definition of functional models can be given as follows:

a) The structure of the model is given by a directed acyclic graph.

b) To each node of the graph corresponds a variable. Variables with indegree zero are called exogenous variables and marked by two dots. All other variables are called endogenous variables and marked by a single dot. (The notation corresponds to the convention introduced in **2.1.8**.)

c) For each endogenous variable, there is a deterministic or probabilistic function showing how the variable (its values or probability distribution) depends on values of the immediately preceding variables.

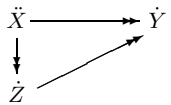
d) Without further assumptions, exogenous variables do not have an associated distribution.

This is a formal framework. The arrows between variables have no specified meaning. In many applications they can be understood as indicating some kind of dependence relation.

3. Illustration with a simple example

To illustrate the notion of functional model, I begin with a simple example which concerns the educational outcome of a generic child.

Model 2.3.1



For simplicity, the variables are assumed to be binary and defined as follows:

$\dot{Y}$	child's educational outcome (1 successful, 0 otherwise)
$\ddot{X}$	parents' educational level (1 high, 0 low)
$\dot{Z}$	school type (1 or 2)

The model contains two probabilistic functions:

$$x \longrightarrow \Pr[\dot{Z} \mid \ddot{X}{=}x]$$
$$(x,z) \longrightarrow \Pr[\dot{Y} \mid \ddot{X}{=}x, \dot{Z}{=}z]$$

Since $\dot{Z}$ and $\dot{Y}$ are binary variables, it suffices to consider the functions

$$x \longrightarrow \Pr(\dot{Z}{=}1 \mid \ddot{X}{=}x)$$
$$(x,z) \longrightarrow \Pr(\dot{Y}{=}1 \mid \ddot{X}{=}x, \dot{Z}{=}z)$$

The first function is intended to show how the probability of attending a specified school type depends on the parents' educational level. The second function is intended to show how the probability of educational success depends both on the parents' educational level and on the school type.

4. Assuming distributions for exogenous variables?

Exogenous variables of a functional model do not have an associated distribution. However, there sometimes are reasons for assuming distributions for exogenous variables.

a) Using the model for a statistical explanation (as defined in **2.1.9**) w.r.t. a particular reference set. Distributions of the model's exogenous variables are then identified with the distributions of the corresponding statistical variables.

b) Using the model for predicting the outcome for an individual that is (only) known to belong to a particular reference set. One then employs a reduced model that is derived from the original model by integrating over the distributions of the unobserved exogenous variables. Assume, for example, that one wants to use the model of the previous subsection for predicting a child's educational success. Not knowing the educational level of the parents, one cannot use the model. However, substituting $\ddot{X}$ by a variable $\dot{X}$ with a probability distribution $\Pr[\dot{X}|\ddot{Z}{=}z]$, one can derive a *reduced model*

$$\Pr(\dot{Y}{=}1 \mid \ddot{Z}{=}z) = \sum\nolimits_x \Pr(\dot{Y}{=}1 \mid \ddot{Z}{=}z, \dot{X}{=}x) \Pr(\dot{X}{=}x|\ddot{Z}{=}z)$$

which only requires knowledge of the child's school type.

c) Using the model for predicting the value of an exogenous variable based on knowing values of endogenous variables. In order to apply Bayesian inference, one must begin with a prior distribution for the exogenous variables.

5. Defining effects of explanatory variables

Assume that $\dot{Y}$ depends on an exogenous variable $\ddot{X}$. To think of an effect of $\ddot{X}$ means to compare

$$\Pr[\dot{Y}|\ddot{X}{=}x'] \quad \text{and} \quad \Pr[\dot{Y}|\ddot{X}{=}x'']$$

for (at least) two values, x' and x'', of $\ddot{X}$. This comparison concerns conditional distributions and cannot, in general, be summarized by a single number.

One therefore often uses a simplified definition which only compares expected values:

$$\mathrm{E}(\dot{Y}|\ddot{X}{=}x'') - \mathrm{E}(\dot{Y}|\ddot{X}{=}x') \tag{2.3}$$

However, one has to take into account that $\dot{Y}$ also depends on further variables. Then, in general, effects cannot simply be attributed to a change in $\ddot{X}$, but are *context-dependent*. Formally, assume that $\dot{Y}$ also depends on $\ddot{Z}$. The effect of a change in $\ddot{X}$ must then be written as

$$\mathrm{E}(\dot{Y}|\ddot{X}{=}x'', \ddot{Z}{=}z) - \mathrm{E}(\dot{Y}|\ddot{X}{=}x', \ddot{Z}{=}z) \tag{2.4}$$

and, in general, depends on the *covariate context* specified by $\ddot{Z}{=}z$.

6. Explained variance and statistical explanation

Authors who estimate regression functions often report some measure of 'explained variance'. It is noteworthy that this notion cannot immediately be applied to functional models. Assume a simple functional model: $\ddot{X} \longrightarrow\!\!\!\!\!\rightarrow \dot{Y}$. Since there is no distribution for $\ddot{X}$, also $\dot{Y}$ has no distribution, and the idea of 'explained variation' cannot be applied.

It would be possible to refer to the variance of $\dot{Y}$ conditional on values of $\ddot{X}$: $\mathrm{V}(\dot{Y}|\ddot{X}{=}x)$. This could be used for quantifying the uncertainty of predictions; but this is a different idea.

However, the notion of 'explained variance' can sensibly be used when functional models are considered as tools for statistical explanations. As proposed in **2.1.9**, this application starts from a statistical variable, say

$$(X,Y) : \Omega \longrightarrow \mathcal{X} \times \mathcal{Y}$$

One then uses $\mathrm{P}[X]$, the distribution of X, and a functional model $\ddot{X} \longrightarrow\!\!\!\!\!\rightarrow \dot{Y}$ to construct a statistical variable, say $\hat{Y}$, whose distribution approximates $\mathrm{P}[Y]$. Two approaches to the construction of $\mathrm{P}[\hat{Y}]$ can be distinguished:

a) Using a prediction rule for individual outcomes to define individual values of $\hat{Y}$, e.g. $\hat{Y}(\omega) := \mathrm{E}(\dot{Y} \mid \ddot{X} = X(\omega))$.

b) Directly deriving a distribution of $\hat{Y}$.

The first approach depends on specifying a prediction rule which can be done in several different ways, in particular when the outcome variable is qualitative. I therefore focus on the second approach which does not require prediction rules for individual outcomes. Following this approach, the construction of $\hat{Y}$'s distribution begins with conditional values:

$$\mathrm{P}(\hat{Y}{=}y \,|\, X{=}x) := \Pr(\dot{Y} = y \,|\, \ddot{X}{=}x)$$

One then uses the known distribution of X to derive the corresponding distribution of $\hat{Y}$:

$$\mathrm{P}(\hat{Y}{=}y) = \sum\nolimits_x \mathrm{P}(\hat{Y}{=}y \,|\, X{=}x)\,\mathrm{P}(X{=}x)$$

Finally, the construction can be assessed with two considerations:

– One can compare $\mathrm{P}[\hat{Y}]$ with $\mathrm{P}[Y]$ ('goodness of distributional fit').

– One can calculate the part of the variation of $\hat{Y}$ which can be attributed to variation of X ('explained variance').[1]

7. Numerical illustration

To illustrate the construction, I use Model 2.3.1 and assume the following data:

X	Z	$Y{=}0$	$Y{=}1$
0	1	300	300
0	2	80	320
1	1	40	160
1	2	80	720

The goodness of distributional fit depends on the parametric model that is used to approximate the functional model. Using a saturated model, the fit would be perfect. In the example:

$$\Pr(\dot{Y}{=}1|\ddot{X}{=}x, \dot{Z}{=}z){=}\mathrm{P}(Y{=}1|X{=}x, Z{=}z) \implies \mathrm{P}[\hat{Y}]{=}\mathrm{P}[Y]$$

The fit would not be perfect if one had used, for example, a logit model without an interaction effect. And, of course, there will be no perfect fit when the statistical explanation concerns a set of data different from those that are used to estimate the model.

Finally, one can calculate the explained variance, that is, the part of the variance of $\hat{Y}$ (not of Y) which can be attributed to variation of X.

[1] Of cause, one could use other measures of variation instead of variance.

Since the joint distribution of X and $\hat{Y}$ is known, one can apply a standard variance decomposition:

$$\mathrm{V}(\hat{Y}) = \mathrm{V}[\mathrm{M}(\hat{Y}|X)] + \mathrm{M}[\mathrm{V}(\hat{Y}|X)]$$

The first part can be interpreted as *explained variation:*

$$\mathrm{V}[\mathrm{M}(\hat{Y}|X)] = \sum\nolimits_x [\mathrm{M}(\hat{Y}|X{=}x) - \mathrm{M}(\hat{Y})]^2\,\mathrm{P}(X{=}x)$$

The second part is the *residual variation*:

$$\mathrm{M}[\mathrm{V}(\hat{Y}|X)] = \sum\nolimits_x \mathrm{V}(\hat{Y}|X{=}x)\,\mathrm{P}(X{=}x)$$

In our illustration, assuming a saturated model, one finds $\mathrm{M}(\hat{Y}) = \mathrm{M}(Y) = 0.75$ and $\mathrm{V}(\hat{Y}) = \mathrm{V}(Y) = 0.1875$, and finally:

– explained variation: $\mathrm{V}[\mathrm{M}(\hat{Y}|X, Z)] = 0.0285$,

– residual variation: $\mathrm{M}[\mathrm{V}(\hat{Y}|X, Z)] = 0.1590$,

– proportion of explained variation: $0.0285/0.1875 \approx 15\%$.

Chapter 3

Estimating effects with logit models

Functional models as introduced in the previous chapter are primarily tools for the formulation of hypotheses about relationships between variables. In order to estimate the hypothesized relations one uses data and, most often, parametric models. One of these parametric models, which is often used when the outcome variable of interest is binary, is the logit model. In the present chapter, I focus on the question of how to understand, and compare, effects of explanatory variables estimated with such models.

3.1 Defining effects with logit models

1. A single explanatory variable

The most simple logit model corresponds to a functional model $\ddot{X} \longrightarrow\!\!\!\!\!\rightarrow \dot{Y}$. $\ddot{X}$ is the explanatory variable, $\dot{Y}$ is the binary outcome variable. As an example, already introduced in Section 2.3, one can think that the model concerns the dependence of children's success in school ($\dot{Y} = 1$ if success, $\dot{Y} = 0$ otherwise) on parents' educational level represented by $\ddot{X}$ (e.g. 0 low, 1 high). The functional model posits a functional relationship

$$x \longrightarrow \Pr(\dot{Y}=1|\ddot{X}=x) = \mathrm{E}(\dot{Y}|\ddot{X}=x) \tag{3.1}$$

This function shows how the expectation of $\dot{Y}$ depends on values of $\ddot{X}$. In the present chapter, the interest concerns effects, that is, effects of *changes (differences) of values* of $\ddot{X}$ on the distribution of $\dot{Y}$. I use the notation

$$\Delta^s(\dot{Y}; \ddot{X}[x', x'']) := \mathrm{E}(\dot{Y}|\ddot{X}=x'') - \mathrm{E}(\dot{Y}|\ddot{X}=x') \tag{3.2}$$

and refer to this as the stochastic (in contrast to: deterministic) effect of a change in the variable $\ddot{X}$ from x' to x''.

The logit model assumes a specific parametric representation of the functional relationship (3.1). It is based on using a logistic link function

$$F(v) := \frac{\exp(v)}{1+\exp(v)} \tag{3.3}$$

to approximate (3.1), resulting in the model

$$\mathrm{E}(\dot{Y}|\ddot{X}=x) \approx F(\alpha + x\beta_x) \tag{3.4}$$

Using here an equality sign instead of $\approx$ would presuppose that the model is 'correctly specified'. However, in particular when thinking of the possibility that further explanatory variables should be included, this cannot be assumed just from the beginning.

The effect of a change in $\ddot{X}$ from x' to x'', as defined in (3.2), is then approximated by

$$\Delta^a(\dot{Y}; \ddot{X}[x', x'']) := F(\alpha + x''\beta_x) - F(\alpha + x'\beta_x) \tag{3.5}$$

where the 'a' is intended to indicate 'approximation'.

2. Adding another explanatory variable

I now consider the addition of another explanatory variable, say $\ddot{Z}$. To continue with the example, one can imagine that the child's success ($\dot{Y}$) not only depends on the parents' educational level ($\ddot{X}$), but also on the school type ($\ddot{Z}$). Graphically depicted, the model then is $(\ddot{X}, \ddot{Z}) \longrightarrow\!\!\!\!\!\rightarrow \dot{Y}$, and the corresponding functional relationship is

$$(x, z) \longrightarrow \Pr(\dot{Y}=1|\ddot{X}=x, \ddot{Z}=z) = \mathrm{E}(\dot{Y}|\ddot{X}=x, \ddot{Z}=z) \tag{3.6}$$

In contrast to the simple model (3.1), effects of $\ddot{X}$ can now be defined only conditional on values of $\ddot{Z}$:

$$\begin{aligned}\Delta^s(\dot{Y}; \ddot{X}[x', x''], \ddot{Z}=z) :=& \\ \mathrm{E}(\dot{Y}|\ddot{X}=x'', \ddot{Z}=z) - \mathrm{E}(\dot{Y}|\ddot{X}=x', \ddot{Z}=z)&\end{aligned} \tag{3.7}$$

Again, one can use a logit model as a parametric approximation to (3.6). Including an interaction term, the model is

$$\mathrm{E}(\dot{Y}|\ddot{X}=x, \ddot{Z}=z) \approx F(\alpha^* + x\beta_x^* + z\beta_z^* + xz\beta_{xz}^*) \tag{3.8}$$

Of course, since this model differs from (3.4), also the parameters must be distinguished. The parameterized effect then is

$$\begin{aligned}\Delta^a(\dot{Y}; \ddot{X}[x', x''], \ddot{Z}=z) :=& \\ F(\alpha^* + x''\beta_x^* + z\beta_z^* + x''z\beta_{xz}^*) - F(\alpha^* + x'\beta_x^* + z\beta_z^* + x'z\beta_{xz}^*)&\end{aligned} \tag{3.9}$$

Table 3.1 Fictitious data for the illustration.

x	z	y	cases
0	0	0	600
0	0	1	600
0	1	0	240
0	1	1	560
1	0	0	40
1	0	1	160
1	1	0	80
1	1	1	720

To illustrate, I use the data shown in Table 3.1. Y represents the child's success ($Y = 1$), X represents the parents' educational level (0 low, 1 high), and Z represents the school type (0 or 1). Nonparametric estimates can be derived directly from the observed frequencies as shown in the following table:

x	z	$\mathrm{E}(Y\|X=x, Z=z)$
0	0	0.5
0	1	0.7
1	0	0.8
1	1	0.9

(3.10)

(Using the logit model (3.8) would result in identical estimates. Leaving out the interaction effect would lead to slightly different values.) One then finds the effects:

$$\Delta^s(\dot{Y}; \ddot{X}[0,1], \ddot{Z}=0) = 0.8 - 0.5 = 0.3 \qquad (3.11)$$
$$\Delta^s(\dot{Y}; \ddot{X}[0,1], \ddot{Z}=1) = 0.9 - 0.7 = 0.2$$

showing that there is no unique effect, but that the effect of the parents' educational level depends on the school type.

3. Parameters in reduced models

The parameters β_x and β_x^* cannot immediately be compared and must be considered as belonging to different models. In order to stress this point, I briefly criticize the idea that parameters in reduced models should be viewed as 'biased estimates' of corresponding parameters in more comprehensive models. To illustrate the argument, I use an example taken from Mood (2010: 71). The example assumes a correctly specified logit model

$$\mathrm{E}(\dot{Y}|\ddot{X}=x, \ddot{Z}=z) = F(x\beta_x + z\beta_z) \qquad (3.12)$$

Values of $\ddot{X}$ and $\ddot{Z}$ are taken from two independent standard normal distributions. Written with a latent variable, the model is

$$\dot{Y}_l := x\beta_x + z\beta_z + \dot{L} \qquad (3.13)$$

where $\dot{L}$ is a random variable with a standard logistic distribution, defined by $\Pr(\dot{L} \le l) = F(l)$, implying that $\dot{Y}_l \ge 0 \iff \dot{Y} = 1$ (based on the symmetry of the distribution of $\dot{L}$).[1] Mood uses this model with $\beta_x = 1$ and three different values for β_z. I begin with assuming that also $\beta_z = 1$.

One can then consider a model which omits $\ddot{Z}$. Taken as a standard logit model, it can be written in terms of a latent variable as

$$\dot{Y}_l^r := x\beta_x^r + \dot{L} \qquad (3.14)$$

Estimating this model with simulated data, Mood finds $\beta_x^r = 0.84$, which is obviously less than $\beta_x = 1$, and concludes that the estimate is 'clearly biased towards zero' (p. 71).[2] However, this statement presupposes that (3.14) has the task to estimate β_x as defined by (3.12), and this is at least debatable.

Viewing (3.14) as a reduced version of (3.12), it provides estimates of probabilities which have a clear and sensible meaning: they approximate probabilities which are averages w.r.t. the (a presupposed) distribution of the omitted variable. In the example, $F(x\beta_x^r)$ approximates

$$\mathrm{E}_{\dot{Z}}(\Pr(\dot{Y}=1|\ddot{X}=x, \dot{Z})) := \int_z F(x\beta_x + z\beta_z)\,\phi(z)\,dz \qquad (3.15)$$

where $\phi(z)$ denotes the standard normal density function. This shows that β_x^r is the correct parameter to be used when being interested in approximating the probabilities defined in (3.15). Instead intending to estimate β_x would not be sensible. In fact, knowing β_x without also knowing β_z would be almost useless because $F(x\beta_x)$ provides a correct estimate only for the special case where $z = 0$.

Note that the proposed interpretation of the reduced model (3.14) holds independently of the size of β_z. For example, assuming $\beta_z = 2$, Mood finds $\beta_x^r = 0.61$, even smaller than 0.84, but $F(x\beta_x^r)$ is still an (actually very good) approximation to the average w.r.t. the omitted variable as defined in (3.15).

[1] It is often said that the variance of the latent variable $\dot{Y}_l$ is 'not identified' (e.g., Allison 1999, Cramer 2007). This is true in the following sense: When starting from a regression model $\dot{Y}_l = x\beta_x + z\beta_z + \epsilon$ with an arbitrary residual variable ϵ, and observations (values of $\dot{Y}$) only provide information about the sign of $\dot{Y}_l$, the variance of this variable cannot be estimated. The statement is misleading, however, when the latent variable is derived from a logit model. If $\dot{Y}_l$ is *defined* by (3.13), a variance of $\dot{Y}_l$ does exist only conditional on values of the explanatory variables, and is already known from the model's definition: $\mathrm{Var}(\dot{Y}_l|\ddot{X}=x, \ddot{Z}=z) = \mathrm{Var}(\dot{L}) = \pi^2/3$.

[2] For similar views see Allison (1999), Cramer (2007), Wooldridge (2002: 470).

3.2 Comparing effects across models

I now consider the question of how to compare the effects of $\ddot{X}$ across the two models, (3.1) and (3.6).

1. Consideration of composite effects

Obviously, an immediate comparison is not possible because in model (3.6) effects also depend on values of $\ddot{Z}$. One therefore needs to define *composite effects* based on a reduced version of (3.6). This requires to think of $\ddot{Z}$ as a random variable, say $\dot{Z}$, that has an associated distribution. Taking into account that the distribution of $\dot{Z}$ could depend on values of $\ddot{X}$, one can start from the equation

$$\mathrm{E}(\dot{Y}|\ddot{X}=x) = \sum\nolimits_z \mathrm{E}(\dot{Y}|\ddot{X}=x, \dot{Z}=z)\,\Pr(\dot{Z}=z|\ddot{X}=x) \tag{3.16}$$

Here I assume that $\dot{Z}$ is a discrete variable as it is the case in the school example; if $\dot{Z}$ is continuous, as it is the case in Mood's example, one would use an integral instead of the sum. The effect defined in (3.2) can then be expressed as

$$\begin{aligned}\Delta^s(\dot{Y}; \ddot{X}[x', x'']) = &\sum\nolimits_z \mathrm{E}(\dot{Y}|\ddot{X}=x'', \dot{Z}=z)\,\Pr(\dot{Z}=z|\ddot{X}=x'') \\ &- \sum\nolimits_z \mathrm{E}(\dot{Y}|\ddot{X}=x', \dot{Z}=z)\,\Pr(\dot{Z}=z|\ddot{X}=x')\end{aligned} \tag{3.17}$$

A simpler formulation is possible if $\dot{Z}$ is independent of $\ddot{X}$. The composite effect is then an average of the conditional effects:

$$\Delta^s(\dot{Y}; \ddot{X}[x', x'']) = \sum\nolimits_z \Delta^s(\dot{Y}; \ddot{X}[x', x''], \dot{Z}=z)\,\Pr(\dot{Z}=z) \tag{3.18}$$

Note, however, that even in this case the effect of $\ddot{X}$ depends on the distribution of $\dot{Z}$ (presupposing that the effect of $\ddot{X}$ depends on $\dot{Z}$).

To illustrate, I use Mood's example where $\dot{Z}$ has a normal distribution independent of $\ddot{X}$. Corresponding to (3.18) one finds the approximation

$$\Delta^s(\dot{Y}; \ddot{X}[x', x'']) \approx \int_z (F(x''\beta_x + z\beta_z) - F(x'\beta_x + z\beta_z))\,\phi(z)\,dz$$

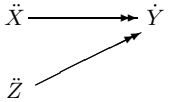
showing how effects of $\ddot{X}$ also depend on the distribution of $\dot{Z}$. For example, assuming $\dot{Z} \sim \mathcal{N}(0,1)$, one finds $\Delta^s(\dot{Y}; \ddot{X}[0,1]) \approx 0.7 - 0.5 = 0.2$, but the effect will increase when the variance of $\dot{Z}$ becomes smaller and, conversely, will decrease when the variance becomes larger.

2. Correlated explanatory variables

In social research, explanatory variables are most often correlated, and the simple relationship (3.18) does not hold. A first problem then concerns how to think of correlations between observed explanatory variables. A further problem that will be deferred to a later section concerns possibly relevant omitted variables which, presumably, are correlated with already included explanatory variables.

How to take into account correlations between observed explanatory variables depends on the purpose of the model to be estimated. One purpose of a model could be to describe the relationship between a dependent and several explanatory variables as found in a given data set (and assumed to exist in a correspondingly defined population). Given this purpose, one can ignore correlations between explanatory variables and, assuming two such variables, refer to a model as follows:

Model 3.2.1



The model only concerns the dependency of the probability distribution of $\dot{Y}$ on values of the two explanatory variables and does not entail anything about relationships between these variables. In other words, the explanatory variables are treated as exogenous variables without associated distributions; and this entails that the model cannot be used to think about correlations between these variables. Of course, the model can be estimated also with data exhibiting correlations between the explanatory variables. Think for example of the data in Table 3.1 where the statistical variables corresponding to $\ddot{X}$ and $\ddot{Z}$ are correlated.

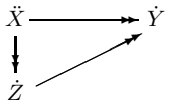
Another purpose of a model could be to investigate effects of variables as defined in the first section. For example, one might be interested in the question of how the expectation of $\dot{Y}$ (the child's success) depends on a change, or difference, in the variable $\ddot{X}$ (the parents' educational level). Obviously, Model 3.2.1 cannot be used to answer this question because the effect also depends on values of $\ddot{Z}$. The observation of correlations between explanatory variables then leads to an important question: Can values of $\ddot{Z}$ be fixed when referring to the effect of a change in the value of $\ddot{X}$?

Of course, given a function like (3.6), one can easily think of changes in values of $\ddot{X}$, and consequently of effects of $\ddot{X}$, while holding $\ddot{Z} = z$ fixed. However, in a more relevant understanding the question does not concern possibilities to manipulate formulas, but the behavior of the social processes which actually generate values of the variables represented in a model (see Rohwer 2010: 82ff). In this understanding, the question motivates to consider more comprehensive models which include assumptions about relationships between explanatory variables.

3. Modeling dependency relations

There are several different possibilities. Here I briefly consider two (further possibilities will be considered in the next chapter). The first one can be depicted as follows:

Model 3.2.2



$\ddot{X}$ is still an exogenous variable, but $\ddot{Z}$ has now changed into an endogenous stochastic variable, $\dot{Z}$. In addition to the function (3.6), there is now another function

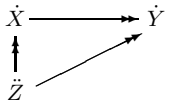
$$x \longrightarrow \Pr(\dot{Z}=z|\ddot{X}=x) \tag{3.19}$$

showing how the distribution of $\dot{Z}$ depends on values of $\ddot{X}$. In our example, based on the data in Table 3.1, one finds $\Pr(\dot{Z}=1|\ddot{X}=0)=0.4$ and $\Pr(\dot{Z}=1|\ddot{X}=1)=0.8$, showing how the child's school type depends on the parents' educational level.

Given this model, a change in $\ddot{X}$ entails a change in the distribution of $\dot{Z}$ and it is not possible to fix $\dot{Z}=z$ when considering an effect of $\ddot{X}$. Consequently, when considering this model, one can only define a total effect of a change in $\ddot{X}$, and this total effect equals the effect (3.17) which is derived from a reduced model resulting from omitting $\dot{Z}$; in the example: $\Delta^s(\dot{Y};\ddot{X}[0,1]) \approx 0.88-0.58=0.3$. In other words, assuming Model 3.2.2, marginalization w.r.t. $\dot{Z}$ is required in order to define the effect of interest.

The situation is less clear when considering a model in which the explanatory variable of interest is endogenous, for example:

Model 3.2.3



While the model can well be used to define an effect of $\ddot{Z}$, there is no straightforward answer to the question of how to define an effect of a change in $\dot{X}$. One could fix $\ddot{Z}=z$ and nevertheless think of different values of $\dot{X}$ to be used for the calculation of an effect; but such effects are conditional on $\ddot{Z}=z$ and already available in Model 3.2.1. On the other hand, without a distribution for $\ddot{Z}$, one cannot derive a reduced model. Thinking instead of a variable $\dot{Z}$ that can be assumed to have a distribution, the composite

Table 3.2 Modification of the data in Table 3.1.

x	z	y	cases
0	0	0	257
0	0	1	257
0	1	0	360
0	1	1	840
1	0	0	17
1	0	1	69
1	1	0	120
1	1	1	1080

effect of $\dot{X}$ depends on the actual choice. For example, deriving the distribution of $\dot{Z}$ from the data in Table 3.1, one finds the composite effect 0.3. Using instead the data in Table 3.2 (which entail the same functional relationships as specified in (3.10)), one finds 0.25. Given this model, it seems best not to attempt to attribute to $\dot{X}$ a definite (context-independent) effect.

4. Continuing with Mood's example

For further illustration of correlated explanatory variables, I use a modification of Mood's example in which values of $\ddot{X}$ and $\ddot{Z}$ are taken from a bivariate normal distribution with correlation $\rho \neq 0$. One can again consider the reduced model (3.14). For example, assuming $\rho=0.5$, one finds $\beta_x^r=1.32$, now larger than $\beta_x=1$ (this also shows that omitting a variable not always leads to an 'attenuated parameter'). As I have argued above, this is not a 'biased estimate' of β_x, but must be viewed as a parameter of the reduced model (3.14). In this understanding, β_x^r can be used to calculate a sensible approximation to the expectation (3.16). In the example, $\mathrm{E}(\dot{Y}|\ddot{X}=0) \approx F(0)=0.5$, and $\mathrm{E}(\dot{Y}|\ddot{X}=1) \approx F(1.32)=0.79$.

These values could be used to calculate the effect $\Delta^s(\dot{Y};\ddot{X}[0,1]) \approx 0.79-0.5=0.29$, obviously larger than the value 0.2 that was calculated for Mood's original model with uncorrelated explanatory variables. In order to understand the difference, one needs an extended model that allows one to interpret the correlation between the two explanatory variables. I consider Model 3.2.2 which is based on the assumption that the distribution of $\dot{Z}$ depends on values of $\ddot{X}$. In the example, the conditional density of $\dot{Z}$, given $\ddot{X}=x$, is a normal density $\phi(z;\mu,\sigma)$ with $\mu=x\rho$ and $\sigma=\sqrt{1-\rho^2}$, entailing that $\ddot{X}$ and $\dot{Z}$ are connected by a linear regression function.

This allows an easy interpretation of the effect. For example, if the value of $\ddot{X}$ changes from 0 to 1, this entails a change in the mean value of $\dot{Z}$ from 0 to ρ, and, if $\rho>0$, the effect becomes larger compared with a

Table 3.3 Modification of the data in Table 3.1.

x	z	y	cases
0	0	0	400
0	0	1	600
0	1	0	100
0	1	1	400
1	0	0	200
1	0	1	300
1	1	0	200
1	1	1	800

situation where $\rho = 0$. In any case, assuming that $\dot{Z}$ depends on $\ddot{X}$ allows one to attribute the total effect to the change in $\ddot{X}$.

5. Comparing variables across models

Neither parameters nor effects can directly be compared across models. It is well possible, however, to compare the role played by explanatory variables. For example, one can compare the role played by $\ddot{X}$ across the models (3.1) and (3.6). One can begin with a look at the estimated parameters. Using the data in Table 3.1, one finds $\hat{\beta}_x = 1.67$ and $\hat{\beta}^*_x = 1.39$. This does not show, however, that $\ddot{X}$ is 'less important' when one 'controls for' values of $\ddot{Z}$. The total effect of $\ddot{X}$ is essentially identical in both models (differences only result from the parameterization of the models). Of course, the enlarged model provides an opportunity to think of this total effect in a more refined way.

Even if, by including a further variable, a parameter becomes zero one cannot conclude that the corresponding variable has no effect. To illustrate, I use the data in Table 3.3. Using these data to estimate (3.4) and (3.8), one finds $\hat{\beta}_x = 0.32$ and $\hat{\beta}^*_x = 0$. This shows that the effect of $\ddot{X}$, conditional on values of $\ddot{Z}$, is zero. There nevertheless is a relevant total effect of $\ddot{X}$, namely $\Delta^s(\dot{Y}; \ddot{X}[0,1]) \approx 0.73 - 0.67 = 0.06$.

How to interpret this effect depends on assumptions about the relationship between $\ddot{X}$ and $\ddot{Z}$. In our example, assuming that the choice of a school type depends on the parents' educational level, one would use Model 3.2.2. The total effect of $\ddot{X}$ can then be explained by the difference in the probabilities $\Pr(\dot{Z}=1|\ddot{X}=0) = 1/3$ and $\Pr(\dot{Z}=1|\ddot{X}=1) = 2/3$.

6. Unobserved heterogeneity

So far, I have assumed observed explanatory variables. Further questions concern 'unobserved heterogeneity'. I take this expression to mean that there are further unobserved explanatory variables that should be included

in a model. So the question arises how the model would change if these additional variables would have been included. A reliable answer is obviously not possible, but a few remarks can be derived from the foregoing discussion.

As before, I only consider logit models and begin with assuming that the interest concerns conditional expectations,

$$\mathrm{E}(\dot{Y}|\ddot{X}=x) \approx F(\alpha + x\beta_x) \tag{3.20}$$

When hypothetically adding a further explanatory variable, say $\ddot{Z}$, one gets a more comprehensive model. However, in order to think of (3.20) as a reduced version of that model, one needs to think of $\ddot{Z}$ as a variable $\dot{Z}$ that can be assumed to have a distribution (given, e.g., by values of $\ddot{Z}$ if such values could be observed). Equation (3.16) then shows that $\mathrm{E}(\dot{Y}|\ddot{X}=x)$ can be viewed as a mean value w.r.t. the distribution of $\dot{Z}$; and consequently $F(\alpha + x\beta_x)$ can be viewed as an approximation to this mean value. As an illustration remember Mood's example. Not having observed $\ddot{Z}$, one can estimate only the reduced model (3.14), but this model correctly provides an approximation to the expectation defined in (3.15). As shown by (3.16), this remains true when $\dot{Z}$ is correlated with $\ddot{X}$.

The situation is more complicated when the interest concerns effects as defined in (3.2). First assume that the hypothetically included unobserved variable $\dot{Z}$ is independent of the variable $\ddot{X}$. As shown by (3.18), the effect derived from the reduced model can then be viewed as a mean of effects which additionally condition on value of $\dot{Z}$. Of course, the not observed effects $\Delta^s(\dot{Y}; \ddot{X}[x', x''], \dot{Z} = z)$ can have quite different, even positive and negative, values. For example, one can easily modify the data in Table 3.1 to get conditional expectations as follows:

x	z	$\mathrm{E}(\dot{Y}\|\ddot{X}=x, \ddot{Z}=z)$
0	0	0.7
0	1	0.8
1	0	0.6
1	1	0.9

entailing effects $\Delta^s(\dot{Y}; \ddot{X}[0,1], \ddot{Z}=0) = -0.1$ and $\Delta^s(\dot{Y}; \ddot{X}[0,1], \ddot{Z}=1) = 0.1$. The observed effect is then positive if $\Pr(\dot{Z}=1) > 0.5$ and negative otherwise.

When the omitted variable is correlated with observed explanatory variables, a critical question concerns the sources of the correlation. To conceive of the observed effect of $\ddot{X}$ as a total effect requires the presupposition of a model in which the omitted variables functionally depends on $\ddot{X}$. Otherwise, as I have argued above, no easy interpretation of the observed effect seems possible.

Chapter 4

Relationships between explanatory variables

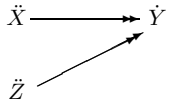
Statistical explanations most often consider many explanatory variables. The formulation of explanatory claims must include, then, an explication of relationships between the explanatory variables. This chapter discusses how functional models can be used to consider such relationships. The first section deals with interactions between explanatory variables and points to implications for understanding effects. The second section considers functional relationships between explanatory variables (which are different from interactions). In both sections, the discussion is general without presupposing a specific parametric form. The third section considers relationships between explanatory variables in the context of linear models for expectations.

4.1 Interactions between explanatory variables

1. A general definition of interaction

The leading idea is: $\ddot{X}$ and $\ddot{Z}$ are *interactive conditions* for the distribution of $\dot{Y}$ if the effect of a change in $\ddot{X}$ $[\ddot{Z}]$ depends on values of $\ddot{Z}$ $[\ddot{X}]$. The formulation shows that the presence of interaction also depends on the definition of 'effect'. To illustrate, I use

Model 4.1.1



with $\dot{Y}$ = indicator of a child's educational success; $\ddot{X}$ = educational level of the child's parents (0 low, 1 high); $\ddot{Z}$ = type of school the child is attending (0 or 1). The corresponding probabilistic function is

$$(x, z) \longrightarrow \Pr[\dot{Y}|\ddot{X}=x, \ddot{Z}=z] \tag{4.1}$$

The right-hand side denotes the probability distribution of $\dot{Y}$ in a situation where $\ddot{X}=x$ and $\ddot{Z}=x$. Assuming that $\dot{Y}$ is discrete, specific values of this distribution will be denoted by $\Pr(\dot{Y}=y|\ddot{X}=x, \ddot{Z}=z)$.

A simple definition of the effect of a change from $\ddot{X}=x'$ to $\ddot{X}=x''$ is given by

$$\Delta^s(\dot{Y}; \ddot{X}[x', x''], \ddot{Z}=z) := \mathrm{E}(\dot{Y}|\ddot{X}=x'', \ddot{Z}=z) - \mathrm{E}(\dot{Y}|\ddot{X}=x', \ddot{Z}=z) \tag{4.2}$$

Using this definition, one can easily invent examples with and without interaction between the two explanatory variables:

with interaction			without interaction		
x	z	$\mathrm{E}(\dot{Y}\|\ddot{X}=x, \ddot{Z}=z)$	x	z	$\mathrm{E}(\dot{Y}\|\ddot{X}=x, \ddot{Z}=z)$
0	0	0.5	0	0	0.5
0	1	0.7	0	1	0.7
1	0	0.8	1	0	0.7
1	1	0.9	1	1	0.9

(4.3)

2. Interaction in parametric models

The above definition of interaction between explanatory variables is independent of the parametric form of the function which relates these variables

to the distribution of an outcome variable. When using parametric models, it depends on the parametric form whether, and how, interactions can be made visible.

Linear models for mean values require an explicit formulation of interaction terms (most often defined by multiplying variables). In contrast, almost all nonlinear models entail interaction effects already by virtue of their mathematical form. However, even then it depends on the details of the parametric form which interaction effects can be made visible.

The logit model can serve as an example. For Model 4.1.1, the standard formulation would be:

$$\Pr(\dot{Y}=1|\ddot{X}=x,\ddot{Z}=z) \approx \frac{\exp(\alpha + x\beta_x + z\beta_z)}{1+\exp(\alpha + x\beta_x + z\beta_z)}$$

This formulation implies an interaction effect when using the effect definition (4.2), but not when using odds ratios:

$$\frac{\Pr(\dot{Y}=1|\ddot{X}=x'',\ddot{Z}=z)/\Pr(\dot{Y}=0|\ddot{X}=x'',\ddot{Z}=z)}{\Pr(\dot{Y}=1|\ddot{X}=x',\ddot{Z}=z)/\Pr(\dot{Y}=0|\ddot{X}=x',\ddot{Z}=z)} \approx \exp\left((x''-x')\,\beta_x\right)$$

Of course, it will often be sensible to add explicitly an interaction term:

$$\Pr(\dot{Y}=1|\ddot{X}=x,\ddot{Z}=z) \approx \frac{\exp(\alpha + x\beta_x + z\beta_z + xz\beta_{xz})}{1+\exp(\alpha + x\beta_x + z\beta_z + xz\beta_{xz})}$$

3. Implications for understanding effects

Interaction has an important consequence: If two explanatory variables interact, no one can be attributed a unique effect. Instead, one must speak of *context-dependent effects*. This notion is symmetrical: each of the interacting variables can be considered as providing a context for the other one.

To illustrate, consider Model 4.1.1 as exemplified on the left-hand side of (4.3). The effect of parents' educational level depends on the school type:

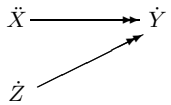
$$\Delta^s(\dot{Y};\ddot{X}[0,1],\ddot{Z}=0) = 0.8-0.5 = 0.3$$
$$\Delta^s(\dot{Y};\ddot{X}[0,1],\ddot{Z}=1) = 0.9-0.7 = 0.2$$

Conversely, the effect of the school type depends on the parents' educational level.

4. Assuming distributions for exogenous variables

In general, whether it is possible to define unique effects also depends on whether the interacting variables are exogenous or endogenous (as specified in a functional model). In Model 4.1.1, both explanatory variables are exogenous. In order to define a unique effect of one of the variables, say $\ddot{X}$, it could be sensible to assume a distribution for the other variable. $\ddot{Z}$ is then substituted by a variable $\dot{Z}$ which is still an exogenous variable with a distribution not depending on $\ddot{X}$. This leads to a modified model:

Model 4.1.2



This modified model would allow one to define a mean effect:

$$\Delta^s(\dot{Y};\ddot{X}[x',x''],\dot{Z}) := \sum\nolimits_z \left(\mathrm{E}(\dot{Y}|\ddot{X}=x'',\dot{Z}=z) - \mathrm{E}(\dot{Y}|\ddot{X}=x',\dot{Z}=z)\right)\Pr(\dot{Z}=z)$$

However, given that $\ddot{X}$ and $\dot{Z}$ interact, the effect still depends on the distribution of $\dot{Z}$. This can easily be seen when using the data on the left-hand side of (4.3):

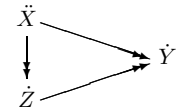
$$\Delta^s(\dot{Y};\ddot{X}[0,1],\dot{Z}) = (0.8-0.5)\Pr(\dot{Z}=0) + (0.9-0.7)\Pr(\dot{Z}=1)$$

4.2 Functional relations between explanatory variables

1. Functional relationships and interactions

The idea of interaction concerns the probabilistic function which relates the explanatory variables to the outcome variable. A different question concerns functional relationships between explanatory variables. As an example, I consider a modification of Model 4.1.1:

Model 4.2.1



This model, in addition to the functional relationship

$$(x,z) \longrightarrow \Pr[\dot{Y}|\ddot{X}=x,\dot{Z}=z] \qquad (4.4)$$

which corresponds to (4.1), also assumes a relationship

$$x \longrightarrow \Pr[\dot{Z}|\ddot{X}=x] \tag{4.5}$$

that shows how the child's school type depends on the parents' educational level. Whether there is interaction between $\ddot{X}$ and $\dot{Z}$ w.r.t. $\dot{Y}$ is completely independent of this function relating $\ddot{X}$ and $\dot{Z}$. For later illustrations, I assume

$$\Pr(\dot{Z}=1\,|\,\ddot{X}=0)=0.4 \quad \text{and} \quad \Pr(\dot{Z}=1\,|\,\ddot{X}=1)=0.8 \tag{4.6}$$

2. Mediator and moderator variables

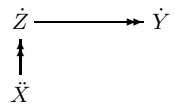
Authors often distinguish between mediator and moderator variables (e.g. Baron and Kenny 1986, MacKinnon 2008). In the context of functional models, one can use the following definitions:

> $\dot{Z}$ is a *mediator variable* for $\ddot{X}$ [or $\dot{X}$] w.r.t. another variable, $\dot{Y}$, if $\dot{Z}$ lies on a directed path leading from $\ddot{X}$ [or $\dot{X}$] to $\dot{Y}$.
>
> $\dot{Z}$ is a *moderator variable* w.r.t. a relationship between $\ddot{X}$ [or $\dot{X}$] and $\dot{Y}$ if the effect of $\ddot{X}$ [or $\dot{X}$] on $\dot{Y}$ depends on values of $\dot{Z}$.

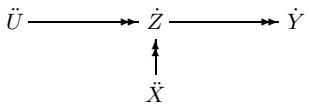
Using these definitions, $\dot{Z}$ is a mediator variable in Model 4.2.1. If $\dot{Z}$ interacts with $\ddot{X}$, it is also a moderator variable. On the other hand, in Model 4.1.1, $\ddot{Z}$ is not a mediator variable, but it is a moderator variable if it interacts with $\ddot{X}$.

One can think, of course, of many different constellations. In the following, I will consider mainly four types:

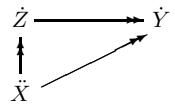
Model 4.2.2a



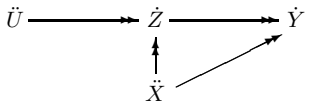
Model 4.2.2c



Model 4.2.2b



Model 4.2.2d



The leading question is, How to define effects of the explanatory variables in these models?

3. Effects of exogenous variables

For models 4.2.2a and 4.2.2b one can derive a *total effect* of $\ddot{X}$, that is, an effect that integrates mediating variables. Starting from

$$\mathrm{E}(\dot{Y}|\ddot{X}=x)=\sum\nolimits_z \mathrm{E}(\dot{Y}|\ddot{X}=x,\dot{Z}=z)\Pr(\dot{Z}=z|\ddot{X}=x)$$

the total effect in 4.2.2b is:

$$\Delta^s(\dot{Y};\ddot{X}[x',x''])=\sum\nolimits_z \Big(\mathrm{E}(\dot{Y}|\ddot{X}=x'',\dot{Z}=z)\Pr(\dot{Z}=z|\ddot{X}=x'')- \mathrm{E}(\dot{Y}|\ddot{X}=x',\dot{Z}=z)\Pr(\dot{Z}=z|\ddot{X}=x')\Big) \tag{4.7}$$

The total effect in 4.2.2a is:

$$\Delta^s(\dot{Y};\ddot{X}[x',x''])= \sum\nolimits_z \mathrm{E}(\dot{Y}|\dot{Z}=z)\left(\Pr(\dot{Z}=z|\ddot{X}=x'')-\Pr(\dot{Z}=z|\ddot{X}=x')\right)$$

To illustrate the calculation of a total effect in Model 4.2.2b, I use the data in (4.3) and (4.6).

Left-hand side of (4.3):

$$\mathrm{E}(\dot{Y}|\ddot{X}=1)=0.8\cdot 0.2+0.9\cdot 0.8=0.88$$
$$\mathrm{E}(\dot{Y}|\ddot{X}=0)=0.5\cdot 0.6+0.7\cdot 0.4=0.58$$
$$\text{Total effect}=0.88-0.58=0.3$$

Right-hand side of (4.3):

$$\mathrm{E}(\dot{Y}|\ddot{X}=1)=0.7\cdot 0.2+0.9\cdot 0.8=0.86$$
$$\mathrm{E}(\dot{Y}|\ddot{X}=0)=0.5\cdot 0.6+0.7\cdot 0.4=0.58$$
$$\text{Total effect}=0.86-0.58=0.28$$

4. Effects of endogenous variables

When thinking of effects of endogenous variables, a first difficulty concerns that such variables have distributions which depend on values of other variables. To circumvent this difficulty, I assume that one can nevertheless hypothetically fix values of endogenous variables. Given this presupposition, one can immediately define effects of $\dot{Z}$ in models 4.2.2a and 4.2.2c:

$$\Delta^s(\dot{Y};\dot{Z}[z',z''])=\mathrm{E}(\dot{Y}|\dot{Z}=z'')-\mathrm{E}(\dot{Y}|\dot{Z}=z')$$

In models 4.2.2b and 4.2.2d, effects of $\dot{Z}$ are context-dependent on values of $\ddot{X}$. For example, in 4.2.2b:

$$\Delta^s(\dot{Y};\dot{Z}[z',z''],\ddot{X}=x)=\mathrm{E}(\dot{Y}|\dot{Z}=z'',\ddot{X}=x)-\mathrm{E}(\dot{Y}|\dot{Z}=z',\ddot{X}=x)$$

5. Direct and indirect effects

Model 4.2.2b (= 4.2.1) leads to the further question of whether one can define, not only a total, but also a *direct effect* of $\ddot{X}$ on the expectation of $\dot{Y}$. A positive answer requires that it makes sense to hypothetically held constant a value of $\dot{Z}$, although its distribution changes with $\ddot{X}$.

Even given this presupposition, one can only define a direct effect if $\ddot{X}$ and $\dot{Z}$ do not interact. If there is no interaction, the effect

$$\Delta^s(\dot{Y}; \ddot{X}[x', x''], \ddot{Z}=z)$$

is independent of z and can sensibly be interpreted as a direct effect of $\ddot{X}$. This is illustrated by the data on the right-hand side of (4.3):

$$\Delta^s(\dot{Y}; \ddot{X}[0,1], \dot{Z}=0) = \Delta^s(\dot{Y}; \ddot{X}[0,1], \dot{Z}=1) = 0.2$$

An indirect effect can then be defined as the difference between the total and the direct effect. Starting from the total effect (4.7), the indirect effect can be written as

$$\sum\nolimits_z \mathrm{E}(\dot{Y}|\ddot{X}=x', \dot{Z}=z)\left(\Pr(\dot{Z}=z|\ddot{X}=x'') - \Pr(\dot{Z}=z|\ddot{X}=x')\right) \qquad (4.8)$$

Using again the right-hand side of (4.3), the indirect effect is 0.08, and the total effect is $0.2 + 0.08 = 0.28$.

If, however, $\ddot{X}$ and $\dot{Z}$ do interact, it is not possible to define a unique direct effect even if one assumes that values of $\dot{Z}$ can be fixed. This is illustrated by the data on the left-hand side of (4.3):

$$\Delta^s(\dot{Y}; \ddot{X}[0,1], \dot{Z}=0) = 0.3 \text{ and } \Delta^s(\dot{Y}; \ddot{X}[0,1], \dot{Z}=1) = 0.2$$

Consequently, there also is no unique indirect effect.

6. Counterfactual effect decompositions

If there is an interaction between $\ddot{X}$ and the mediating variable $\dot{Z}$, direct effects can only be defined for each value of $\dot{Z}$ separately. It is, of course, possible to define versions of mean direct effects. Following this idea, some authors have proposed a 'natural direct effect' (e.g. Pearl 2001, Petersen et al. 2006). In the framework of functional models, this form of a mean direct effect can be defined as

$$\sum\nolimits_z \left(\mathrm{E}(\dot{Y}|\ddot{X}=x'', \dot{Z}=z) - \mathrm{E}(\dot{Y}|\ddot{X}=x', \dot{Z}=z)\right) \Pr(\dot{Z}=z|\ddot{X}=x') \qquad (4.9)$$

The idea is to use the distribution of $\dot{Z}$ that corresponds to the initial value of $\ddot{X}$ and to assume (counterfactually) that this distribution would not change when the value of $\ddot{X}$ changes from x' to x''.

Of course, one also may use other distributions of $\dot{Z}$. In any case, the total effect of $\ddot{X}$ can be divided into a mean direct and a mean indirect effect. Using (4.9) for the mean direct effect, one gets

$$\begin{aligned}\mathrm{E}(\dot{Y}|\ddot{X}=x'') - \mathrm{E}(\dot{Y}|\ddot{X}=x') = \qquad (4.10)\\ \sum\nolimits_z \left[\mathrm{E}(\dot{Y}|\ddot{X}=x'', \dot{Z}=z) - \mathrm{E}(\dot{Y}|\ddot{X}=x', \dot{Z}=z)\right] \Pr(\dot{Z}=z|\ddot{X}=x') + \\ \sum\nolimits_z \mathrm{E}(\dot{Y}|\ddot{X}=x'', \dot{Z}=z)\left[\Pr(\dot{Z}=z|\ddot{X}=x'') - \Pr(\dot{Z}=z|\ddot{X}=x')\right]\end{aligned}$$

where the second term on the right-hand side is then interpreted as a mean indirect effect.

To illustrate the decomposition for a situation with interaction, I use the data on the left-hand side of (4.3) together with (4.6). The mean direct effect is

$$(0.8 - 0.5)\,0.6 + (0.9 - 0.7)\,0.4 = 0.26$$

the mean indirect effect is

$$0.8\,(0.2 - 0.6) + 0.9\,(0.8 - 0.4) = 0.04$$

and the total effect is $0.26 + 0.04 = 0.3$.

In sociological research of education, several authors have used similar definitions to distinguish between primary and secondary family effects on educational outcomes. This will be discussed in Section 6.1.

7. Confounders and independent context variables

There is no unique definition of 'confounding variables' (see, e.g., Weinberg 1993). In the context of functional models, I use the following definition:

> A variable $\dot{X}$ [or $\ddot{X}$] is a confounder w.r.t. a functional dependence of $\dot{Y}$ on $\dot{Z}$ if there is a directed path from $\dot{X}$ to $\dot{Y}$ and (a) there is a directed path from $\dot{X}$ [or $\ddot{X}$] to $\dot{Z}$ (direct confounding), or (b) there is a further variable, say $\dot{U}$ [or $\ddot{U}$], and a directed path leads from $\dot{U}$ to $\dot{Z}$ and from $\dot{U}$ to $\dot{X}$ (indirect confounding).

For example, in Model 4.2.2b, $\ddot{X}$ is a directly confounding variable w.r.t. the dependence of $\dot{Y}$ on $\dot{Z}$. In contrast, $\dot{Z}$ is a mediator variable, not a confounder, w.r.t. the relationship between $\ddot{X}$ and $\dot{Y}$. Notice that the definition presupposes a functional model with directional relationships between variables and cannot be formulated in terms of 'correlation'.

The proposed definition distinguishes confounding variables from independent context variables. Without an arrow from $\ddot{X}$ to $\dot{Z}$ in Model 4.2.2b, $\ddot{X}$ would be an *independent context variable*, not a confounder.

The distinction between these two kinds of covariates concerns possibilities to define effects. Different difficulties show up, in particular, when the variables are not observed. Consider the independent context variable, $\ddot{X}$, in Model 4.1.1. This is a context variable for effects of $\ddot{Z}$ on $\dot{Y}$:

$$\Delta^s(\dot{Y};\ddot{Z}[z',z''],\ddot{X}=x) = \mathrm{E}(\dot{Y}|\ddot{Z}=z'',\ddot{X}=x) - \mathrm{E}(\dot{Y}|\ddot{Z}=z',\ddot{X}=x)$$

If $\ddot{X}$ is not observed, one can think instead of a variable, $\dot{X}$, having an unknown distribution. Since this distribution does not depend on $\ddot{Z}$, the observed effect can be considered as a mean effect w.r.t. the unknown distribution of $\dot{X}$:

$$\Delta^s(\dot{Y};\ddot{Z}[z',z'']) = \sum\nolimits_x \Delta^s(\dot{Y};\ddot{Z}[z',z''],\dot{X}=x)\Pr(\dot{X}=x)$$

Moreover, if there is no interaction between $\ddot{Z}$ and $\dot{X}$, one can attribute the effect uniquely to $\ddot{Z}$.

A similar consideration is not possible for unobserved confounding variables. Consider the confounding variable $\ddot{X}$ in Model 4.2.2b (= 4.2.1). If this variable is not observed, one can instead assume a random variable $\dot{X}$ having an unknown distribution. An effect of $\dot{Z}$ can then be expressed as

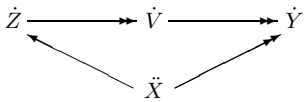
$$\begin{aligned}&\mathrm{E}(\dot{Y}|\dot{Z}=z'') - \mathrm{E}(\dot{Y}|\dot{Z}=z') = \\ &\quad\sum\nolimits_x \Big(\mathrm{E}(\dot{Y}|\dot{Z}=z'',\dot{X}=x)\Pr(\dot{X}=x|\dot{Z}=z'') - \\ &\qquad\qquad \mathrm{E}(\dot{Y}|\dot{Z}=z',\dot{X}=x)\Pr(\dot{X}=x|\dot{Z}=z')\Big)\end{aligned}$$

This effect is not only due to different values of $\dot{Z}$, but also to different distributions of $\dot{X}$ (associated with different values of $\dot{Z}$). The effect cannot, therefore, be interpreted as a mean effect w.r.t. an unknown, but common, distribution of the confounding variable.

8. Effects of mediator variables

Interestingly, the mentioned difficulties do not arise w.r.t. mediator variables. Consider the following

Model 4.2.3



which is similar to Model 4.2.2b but in addition contains the mediator variable $\dot{V}$. As before, if the confounder, $\ddot{X}$, is not observed, the observable relationship between $\dot{Y}$ and $\dot{Z}$ is difficult to interpret. The situation is different, however, w.r.t. the mediator variable $\dot{V}$.

In order to see that, substitute $\ddot{X}$ by a random variable $\dot{X}$ having an unknown distribution. The model entails the relationships

$$\Pr[\dot{Y}|\dot{X}=x,\dot{V}=v,\dot{Z}=z] = \Pr[\dot{Y}|\dot{X}=x,\dot{V}=v] \tag{4.11}$$

$$\Pr[\dot{V}|\dot{Z}=z,\dot{X}=x] = \Pr[\dot{V}|\dot{Z}=z] \tag{4.12}$$

and therefore[1]

$$\Pr[\dot{X}|\dot{V}=v,\dot{Z}=z] = \Pr[\dot{X}|\dot{Z}=z] \tag{4.13}$$

Using these relationships, one finds

$$\begin{aligned}&\mathrm{E}(\dot{Y}|\dot{V}=v,\dot{Z}=z) = \\ &\quad\sum\nolimits_x \mathrm{E}(\dot{Y}|\dot{V}=v,\dot{Z}=z,\dot{X}=x)\Pr(\dot{X}=x|\dot{V}=v,\dot{Z}=z) = \\ &\quad\sum\nolimits_x \mathrm{E}(\dot{Y}|\dot{V}=v,\dot{X}=x)\Pr(\dot{X}=x|\dot{Z}=z)\end{aligned}$$

from which one derives

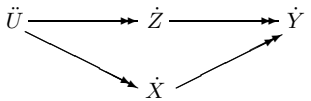
$$\begin{aligned}&\mathrm{E}(\dot{Y}|\dot{V}=v'',\dot{Z}=z) - \mathrm{E}(\dot{Y}|\dot{V}=v',\dot{Z}=z) = \\ &\quad\sum\nolimits_x \big[\mathrm{E}(\dot{Y}|\dot{V}=v'',\dot{X}=x) - \mathrm{E}(\dot{Y}|\dot{V}=v',\dot{X}=x)\big]\Pr(\dot{X}=x|\dot{Z}=z)\end{aligned}$$

This shows that, conditional on values of $\dot{Z}$, effects of $\dot{V}$ can be interpreted as mean effects w.r.t. an unknown distribution of the confounding variable $\dot{X}$. Moreover, if there is no interaction between $\dot{V}$ and $\dot{X}$, one can attribute these effects uniquely to $\dot{V}$.

9. Indirectly connected explanatory variables

So far the discussion dealt with directly confounding variables. I now consider indirectly confounding variables. The following model illustrates this case:

Model 4.2.4



In this model, $\dot{X}$ is a confounder w.r.t. to the dependence of $\dot{Y}$ on $\dot{Z}$ according to the second part of the above definition. If $\dot{X}$ is observed, effects

[1] Starting from

$$\Pr[\dot{X},\dot{V}|\dot{Z}] = \Pr[\dot{X}|\dot{V},\dot{Z}]\Pr[\dot{V}|\dot{Z}] = \Pr[\dot{V}|\dot{X},\dot{Z}]\Pr[\dot{X}|\dot{Z}]$$

and using (4.12), (4.13) immediately follows.

of $\dot{Z}$ can be calculated conditional on values of $\dot{X}$. (Further considerations depend on whether $\dot{X}$ and $\dot{Z}$ interact, see above.)

If $\dot{X}$ is not observed, one can begin with explicitly conditioning on values of $\ddot{U}$ and use the assumption entailed by the model that $\dot{Z}$ and $\dot{X}$ are independent conditional on values of $\ddot{U}$:

$$\begin{aligned}\mathrm{E}(\dot{Y}|\dot{Z}=z,\ddot{U}=u) &= \\ &\sum\nolimits_x \mathrm{E}(\dot{Y}|\dot{Z}=z,\dot{X}=x,\ddot{U}=u)\Pr(\dot{X}=x|\dot{Z}=z,\ddot{U}=u) = \\ &\sum\nolimits_x \mathrm{E}(\dot{Y}|\dot{Z}=z,\dot{X}=x)\Pr(\dot{X}=x|\ddot{U}=u)\end{aligned} \tag{4.14}$$

This shows that, conditional on values of $\ddot{U}$, the observed effect of $\dot{Z}$ can be interpreted as a mean effect w.r.t. an unknown distribution of the confounding variable $\dot{X}$.

Of course, also $\ddot{U}$ is a confounding variable, and no reliable conclusions about effects of $\dot{Z}$ on $\dot{Y}$ can be drawn if also $\ddot{U}$ is not observed.

4.3 Linear regression models

The discussion so far was general and did not rely on any parametric assumptions. I now consider some of the previously discussed relationships in the context of linear models for expectations. (Some considerations for logit models have already been discussed in the previous chapter.)

1. Linear models for expectations

Corresponding to Model 4.1.1, a simple linear version of a parametric model for the expectation of $\dot{Y}$ can be written as

$$\mathrm{E}(\dot{Y}|\ddot{X}=x,\ddot{Z}=z) = \beta_0 + x\beta_x + z\beta_z \tag{4.15}$$

The model entails that $\ddot{X}$ and $\ddot{Z}$ do not interact. Effects of $\ddot{X}$ are then given by

$$\mathrm{E}(\dot{Y}|\ddot{X}=x'') - \mathrm{E}(\dot{Y}|\ddot{X}=x') = (x''-x')\beta_x \tag{4.16}$$

and do not depend on values of $\ddot{Z}$.

What happens if $\ddot{Z}$ is omitted? In order to derive a reduced model, $\ddot{Z}$ must be substituted by a random variable, $\dot{Z}$, having an unknown distribution. If one further assumes that $\dot{Z}$ is independent of $\ddot{X}$, there is a simple result:

$$\mathrm{E}(\dot{Y}|\ddot{X}=x) = \sum\nolimits_z (\beta_0 + x\beta_x + z\beta_z)\Pr(\dot{Z}=z) = \beta_0 + x\beta_x + \mathrm{E}(\dot{Z})\beta_z$$

showing that effects of $\ddot{X}$ are still given by (4.16).

2. Linear models with interactions

In its general formulation, Model 4.1.1 does not make assumptions about interactions between $\ddot{X}$ and $\ddot{Z}$. Such assumptions belong to the form of a parametric model. When using linear models, interaction effects must be explicitly included, in our example:

$$\mathrm{E}(\dot{Y}|\ddot{X}=x,\ddot{Z}=z) = \beta_0 + x\beta_x + z\beta_z + xz\beta_{xz} \tag{4.17}$$

Effects of $\ddot{X}$ now depend on values of $\ddot{Z}$:

$$\mathrm{E}(\dot{Y}|\ddot{X}=x'',\ddot{Z}=z) - \mathrm{E}(\dot{Y}|\ddot{X}=x',\ddot{Z}=z) = (x''-x')(\beta_x + z\beta_{xz})$$

This also changes the consequences of omitting $\ddot{Z}$. If one substitutes $\ddot{Z}$ by $\dot{Z}$ and still assumes that $\dot{Z}$ is independent of $\ddot{X}$, one finds

$$\mathrm{E}(\dot{Y}|\ddot{X}=x) = (\beta_0 + \mathrm{E}(\dot{Z})\beta_z) + x\,(\beta_x + \mathrm{E}(\dot{Z})\beta_{xz})$$

Effects of $\ddot{X}$ now also depend on an unknown mean value of $\dot{Z}$.

3. Mediator variables and total effects

So far I have assumed that $\dot{Z}$ is independent of $\ddot{X}$. I now assume that it is a mediator variable for $\ddot{X}$ w.r.t. $\dot{Y}$, corresponding to Model 4.2.1. Then starting from (4.15) without interaction, one gets

$$\mathrm{E}(\dot{Y}|\ddot{X}=x) = \beta_0 + x\beta_x + \mathrm{E}(\dot{Z}|\ddot{X}=x)\beta_z \tag{4.18}$$

If one assumes a linear model for the expectation of $\dot{Z}$,

$$\mathrm{E}(\dot{Z}|\ddot{X}=x) = \gamma_0 + x\gamma_x \tag{4.19}$$

one also gets a linear formulation for the reduced model:

$$\mathrm{E}(\dot{Y}|\ddot{X}=x) = (\beta_0 + \gamma_0\beta_z) + x\,(\beta_x + \gamma_x\beta_z)$$

Since $\dot{Z}$ is a mediator variable, $\beta_x + \gamma_x\beta_z$ can be interpreted as representing the total effect of $\ddot{X}$ on $\dot{Y}$:

$$\mathrm{E}(\dot{Y}|\ddot{X}=x'') - \mathrm{E}(\dot{Y}|\ddot{X}=x') = (x''-x')(\beta_x + \gamma_x\beta_z)$$

One gets a more involved formulation when starting from (4.17) which includes an interaction effect. Again assuming (4.19), one gets

$$\mathrm{E}(\dot{Y}|\ddot{X}=x) = (\beta_0 + \gamma_0\beta_z) + x\,(\beta_x + \gamma_x\beta_z + \gamma_0\beta_{xz}) + x^2\gamma_x\beta_{xz}$$

$\dot{Y}$ now depends in a nonlinear way on $\ddot{X}$, but $\mathrm{E}(\dot{Y}|\ddot{X}=x'') - \mathrm{E}(\dot{Y}|\ddot{X}=x')$ can still be interpreted as a total effect of $\ddot{X}$ on $\dot{Y}$.

4. Direct and indirect effects

In both cases, the total effect can be decomposed into a direct and an indirect effect. Starting from (4.15) without interaction, one can follow the consideration in **4.2.5**. The direct effect is given by $(x'' - x')\beta_x$, and the indirect effect can be calculated according to (4.8):

$$(\mathrm{E}(\dot{Z}|\ddot{X}=x'') - \mathrm{E}(\dot{Z}|\ddot{X}=x'))\,\beta_z = (x'' - x')\,\gamma_x\beta_z$$

When starting from model (4.17), which includes an interaction effect, one can follow the approach that was described in **4.2.6**. Based on the decomposition (4.10), one finds the counterfactual direct effect

$$(x'' - x')(\beta_x + \gamma_0\beta_{xz} + x'\gamma_x\beta_{xz})$$

and the counterfactual indirect effect

$$(x'' - x')(\gamma_x\beta_z + x''\gamma_x\beta_{xz})$$

5. Omitting a confounding variable

Consider again Model 4.2.1, but now assume that one has omitted the variable $\ddot{X}$. This is a confounding variable w.r.t. the effect of $\dot{Z}$ on $\dot{Y}$. As was argued in **4.2.7**, the effect of $\dot{Z}$ in the reduced model is not easily interpretable. Linear regression models offer no advantage. Starting from (4.15), the reduced model omitting $\dot{X}$ (which is the substitute for $\ddot{X}$) becomes:

$$\mathrm{E}(\dot{Y}|\dot{Z}=z) = \beta_0 + z\beta_z + \mathrm{E}(\dot{X}|\dot{Z}=z)\beta_x$$

While formally similar to (4.18), there is no similar interpretation. In (4.18), one can interpret $\mathrm{E}(\dot{Z}|\ddot{X}=x)\beta_z$ as part of the total effect of $\ddot{X}$ on $\dot{Y}$. In contrast, $\mathrm{E}(\dot{X}|\dot{Z}=z)$ cannot be interpreted as a substantive dependency relation. Starting from the dependency relation assumed in Model 4.2.1, which goes from $\ddot{X}$ to $\dot{Z}$, one can derive

$$\mathrm{E}(\dot{X}|\dot{Z}=z) = \frac{\sum_x x\,\Pr(\dot{Z}=z|\dot{X}=x)\Pr(\dot{X}=x)}{\sum_x \Pr(\dot{Z}=z|\dot{X}=x)\Pr(\dot{X}=x)}$$

So it depends essentially on the distribution of $\dot{X}$ which is determined outside of the model.

Chapter 5

Models of educational stages

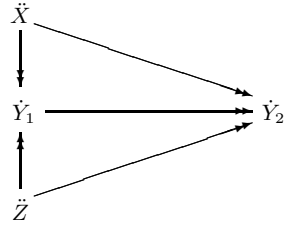
This chapter discusses models of educational processes which consist of two consecutive learning frames, σ_1 and σ_2. The first section considers models of consecutive educational outcomes which presuppose that individuals have values on both variables. The second section considers models of sequential transitions which entail a selection process. (Transition models in which transition probabilities depend on educational outcomes will be discussed in the next chapter.)

5.1 Consecutive educational outcomes

1. Framework of a two-stage model

I consider a sequence of two learning frames, σ_1 and σ_2. Let $\dot{Y}_t$ denote the outcome of learning frame σ_t $(t = 1, 2)$, based on some measure of what has been learned. I assume that these are binary or quantitative variables so that one can sensibly refer to expectations. I consider a model which posits that these outcome variables depend on two explanatory variables, say $\ddot{X}$ (e.g. a measure of socio-economic status) and $\ddot{Z}$ (e.g. an indicator of educational aspiration). The model can be depicted as follows:

Model 5.1



Referring to expectations, the model consists of two stochastic functions:

$$\sigma_1 : \ (x,z) \longrightarrow \mathrm{E}(\dot{Y}_1|\ddot{X}=x,\ddot{Z}=z)$$
$$\sigma_2 : \ (x,z,y) \longrightarrow \mathrm{E}(\dot{Y}_2|\ddot{X}=x,\ddot{Z}=z,\dot{Y}_1=y)$$

The first function is intended to show how the expectation of $\dot{Y}_1$ depends on the two exogenous explanatory variables, the second function is intended to show how the expectation of $\dot{Y}_2$ depends on the same variables and additionally on the outcome of the first learning frame which is used as an endogenous explanatory variable.

2. Conditional and unconditional effects

Without presupposing any particular parametric model, effects of $\ddot{X}$ (SES) on the first outcome can be defined by

$$\Delta^s(\dot{Y}_1;\ddot{X}[x',x''],\ddot{Z}=z) := \mathrm{E}(\dot{Y}_1|\,\ddot{X}=x'',\ddot{Z}=z) - \mathrm{E}(\dot{Y}_1|\,\ddot{X}=x',\ddot{Z}=z) \tag{5.1}$$

Effects of $\ddot{X}$ on the second outcome can be defined in two different ways. One can define *conditional effects* which take into account outcomes of the first learning frame:

$$\Delta^s_c(\dot{Y}_2;\ddot{X}[x',x''],\ddot{Z}=z,\dot{Y}_1=y) := \mathrm{E}(\dot{Y}_2|\,\ddot{X}=x'',\ddot{Z}=z,\dot{Y}_1=y) - \mathrm{E}(\dot{Y}_2|\,\ddot{X}=x',\ddot{Z}=z,\dot{Y}_1=y) \tag{5.2}$$

And one can consider *unconditional effects*:

$$\Delta^s_u(\dot{Y}_2;\ddot{X}[x',x''],\ddot{Z}=z) := \mathrm{E}(\dot{Y}_2|\,\ddot{X}=x'',\ddot{Z}=z) - \mathrm{E}(\dot{Y}_2|\,\ddot{X}=x',\ddot{Z}=z) \tag{5.3}$$

In general, all these effects are context-dependent, that is, depend on values of $\ddot{Z}$. To illustrate, I use the fictitious data in Table 5.1:

Table 5.1 Fictitious data for the two-stage model.

x	z	y_1	y_2	cases	x	z	y_1	y_2	cases
0	0	0	0	300	1	0	0	0	200
0	0	0	1	400	1	0	0	1	300
0	0	1	0	200	1	0	1	0	150
0	0	1	1	600	1	0	1	1	850
0	1	0	0	300	1	1	0	0	200
0	1	0	1	300	1	1	0	1	250
0	1	1	0	200	1	1	1	0	100
0	1	1	1	700	1	1	1	1	950

Effects on first outcome:

$$\Delta^s(\dot{Y}_1;\ddot{X}[0,1],\ddot{Z}=0) = 0.667 - 0.533 = 0.134$$
$$\Delta^s(\dot{Y}_1;\ddot{X}[0,1],\ddot{Z}=1) = 0.700 - 0.600 = 0.100$$

meaning that higher SES has a positive effect on the educational outcome, but the effect is lower for individuals with a high educational aspiration.

Conditionals effect on second outcome:

$$\Delta^s_c(\dot{Y}_2;\ddot{X}[0,1],\ddot{Z}=0,\dot{Y}_1=0) = 0.600 - 0.571 = 0.029$$
$$\Delta^s_c(\dot{Y}_2;\ddot{X}[0,1],\ddot{Z}=1,\dot{Y}_1=0) = 0.556 - 0.500 = 0.056$$
$$\Delta^s_c(\dot{Y}_2;\ddot{X}[0,1],\ddot{Z}=0,\dot{Y}_1=1) = 0.850 - 0.750 = 0.100$$
$$\Delta^s_c(\dot{Y}_2;\ddot{X}[0,1],\ddot{Z}=1,\dot{Y}_1=1) = 0.905 - 0.778 = 0.127$$

Unconditional effect on second outcome:

$$\Delta^s_u(\dot{Y}_2;\ddot{X}[0,1],\ddot{Z}=0) = 0.767 - 0.667 = 0.100$$
$$\Delta^s_u(\dot{Y}_2;\ddot{X}[0,1],\ddot{Z}=1) = 0.800 - 0.667 = 0.133$$

In order to understand the difference between conditional and unconditional effects, it is important to recognize that both concern the outcome in the second learning frame (σ_2). $\dot{Y}_1$ is to be understood as representing a condition for the outcome generation in the second learning frame. This is made explicit in the conditional effect definition. In contrast, the unconditional effect is derived from a reduced model that integrates over the distribution of $\dot{Y}_1$. This is possible, without additional assumptions, because $\dot{Y}_1$ is a mediating variable in the two-stage model:

$$\begin{aligned}\mathrm{E}(\dot{Y}_2|\,\ddot{X}=x,\ddot{Z}=z) = {}& \\ & \mathrm{E}(\dot{Y}_2|\,\ddot{X}=x,\ddot{Z}=z,\dot{Y}_1=0)\Pr(\dot{Y}_1=0|\ddot{X}=x,\ddot{Z}=z) + \\ & \mathrm{E}(\dot{Y}_2|\,\ddot{X}=x,\ddot{Z}=z,\dot{Y}_1=1)\Pr(\dot{Y}_1=1|\ddot{X}=x,\ddot{Z}=z)\end{aligned} \tag{5.4}$$

In this sense, viewing $\dot{Y}_1$ as a mediating variable, the unconditional effect

corresponds to the total effect of $\ddot{X}$ in the context given by $\ddot{Z}$.

3. Effects of endogenous educational outcomes

So far I have considered effects of an exogenous variable, $\ddot{X}$. In formally the same way one could define effects of $\ddot{Z}$ and use $\ddot{X}$ as a context variable. In order to consider effects of the endogenous variable $\dot{Y}_1$, one has to assume that values of both exogenous variables can be fixed. An effect can then be defined by

$$\Delta^s(\dot{Y}_2; \dot{Y}_1[y', y''], \ddot{X}=x, \ddot{Z}=z) := \\ \mathrm{E}(\dot{Y}_2|\,\dot{Y}_1=y'', \ddot{X}=x, \ddot{Z}=z) - \mathrm{E}(\dot{Y}_2|\,\dot{Y}_1=y', \ddot{X}=x, \ddot{Z}=z) \tag{5.5}$$

Like the conditional effects of exogenous variables, this effect concerns the generation of values of $\dot{Y}_2$ in the second learning frame. The effect corresponds to the hypothesis that $\dot{Y}_2$ depends on what has been learned in the preceding learning frame even if the context is fixed by particular values of $\ddot{X}$ and $\ddot{Z}$. With the data in Table 5.1 one finds:

$$\Delta^s(\dot{Y}_2; \dot{Y}_1[0,1], \ddot{X}=0, \ddot{Z}=0) = 0.750 - 0.571 = 0.179$$
$$\Delta^s(\dot{Y}_2; \dot{Y}_1[0,1], \ddot{X}=0, \ddot{Z}=1) = 0.778 - 0.500 = 0.278$$
$$\Delta^s(\dot{Y}_2; \dot{Y}_1[0,1], \ddot{X}=1, \ddot{Z}=0) = 0.850 - 0.600 = 0.250$$
$$\Delta^s(\dot{Y}_2; \dot{Y}_1[0,1], \ddot{X}=1, \ddot{Z}=1) = 0.905 - 0.556 = 0.349$$

showing how the effects of previous learning depend on both $\ddot{X}$ (SES) and $\ddot{Z}$ (aspiration).

4. Endogenously generated inequality

In Model 5.1, one can think that part of the variation of $\dot{Y}_2$ is due to endogenously generated variation of $\dot{Y}_1$, that is, variation conditional on values of the exogenous variables, $\ddot{X}$ and $\ddot{Z}$. Formally, one can use a conditional variance decomposition

$$\mathrm{V}(\dot{Y}_2|x,z) = \mathrm{V}[\mathrm{E}(\dot{Y}_2|\dot{Y}_1,x,z)] + \mathrm{E}[\mathrm{V}(\dot{Y}_2|\dot{Y}_1,x,z)]$$

This is formally analogue to the variance decomposition discussed in **2.3.7**. The first part,

$$\mathrm{V}[\mathrm{E}(\dot{Y}_2|\dot{Y}_1,x,z)] = \\ \sum\nolimits_{y_1} [\mathrm{E}(\dot{Y}_2|\dot{Y}_1=y_1,x,z) - \mathrm{E}(\dot{Y}_2|x,z)]^2 \Pr(\dot{Y}_1=y_1|x,z)$$

is now interpreted as the part of the variation of $\dot{Y}_2$ which is generated by variation of $\dot{Y}_1$. The second part,

$$\mathrm{E}[\mathrm{V}(\dot{Y}_2|\dot{Y}_1,x,z)] = \sum\nolimits_{y_1} \mathrm{V}(\dot{Y}_2|\dot{Y}_1=y_1,x,z) \Pr(\dot{Y}_1=y_1|x,z)$$

is the 'residual variation' which cannot be attributed to $\dot{Y}_1$. Using the data in Table 5.1, one finds the following values.

x	z	$\mathrm{V}(\dot{Y}_2\|x,z)$	$\mathrm{V}[\mathrm{E}(\dot{Y}_2\|\dot{Y}_1,x,z)]$	$\mathrm{E}[\mathrm{V}(\dot{Y}_2\|\dot{Y}_1,x,z)]$
0	0	0.2222	0.0079	0.2143
0	1	0.2222	0.0185	0.2037
1	0	0.1789	0.0139	0.1650
1	1	0.1600	0.0256	0.1344

In this example, the proportion of variation of $\dot{Y}_2$ which is endogenously generated by variation in the educational outcomes of the first learning frame $(\dot{Y}_1|x,z)$, is between 3.5 and 16 %, depending on the values of the exogenous variables.

5. Consequences of omitted variables

I now consider consequences of omitting an explanatory variable. The general strategy is to compare a hypothetically complete model (here I use Model 5.1 for illustrations) with a reduced model that results from integrating over the conditional distribution of the omitted variable. There are then two cases.

The first case occurs if the omitted variable is a mediating variable in the complete model. Its conditional distribution can then be derived from the model without further assumptions. Referring to Model 5.1, this case is illustrated by omitting $\dot{Y}_1$ (see **5.1.2**).

The second case occurs if the conditional distribution of the omitted variable cannot be derived from the complete model without further assumptions. To illustrate, I assume that $\ddot{Z}$ is omitted and one is interested in effects of $\ddot{X}$. In order to derive a reduced model, $\ddot{Z}$ must be substituted by a variable, say $\dot{Z}$, that can be assumed to have a (conditional) distribution. The derivation depends on the kind of effect.

I begin with the effect of $\ddot{X}$ on $\dot{Y}_1$. For the reduced model, one immediately finds

$$\mathrm{E}(\dot{Y}_1|\ddot{X}=x) = \sum\nolimits_z \mathrm{E}(\dot{Y}_1|\ddot{X}=x, \dot{Z}=z) \Pr(\dot{Z}=z|\ddot{X}=x) \tag{5.6}$$

A special case occurs if the effect of $\ddot{X}$ on $\dot{Y}_1$ does not depend on values of $\dot{Z}$. Then, omitting $\dot{Z}$ does not change the relationship between $\ddot{X}$ and $\dot{Y}_1$.[1] In general, the effect of $\ddot{X}$ in the reduced model is a kind of mean value of its context-dependent effects in the complete model.

[1] Note that it is not essential whether $\dot{Z}$ depends on $\ddot{X}$; the essential point concerns whether there is an interaction between $\ddot{X}$ and $\ddot{Z}$ w.r.t. $\dot{Y}_1$.

The same considerations apply to the unconditional effect of $\ddot{X}$ on $\dot{Y}_2$ for which one gets the reduced expression

$$\mathrm{E}(\dot{Y}_2|\ddot{X}=x) = \sum\nolimits_z \mathrm{E}(\dot{Y}_2|\ddot{X}=x, \dot{Z}=z)\,\Pr(\dot{Z}=z|\ddot{X}=x) \tag{5.7}$$

which is completely analogue to (5.6).

When considering the conditional effect of $\ddot{X}$ on $\dot{Y}_2$ one has to take into account that $\ddot{Z}$ is a confounder w.r.t. the relationship between $\dot{Y}_1$ and $\dot{Y}_2$. The reduced relationship is given by

$$\begin{aligned}&\mathrm{E}(\dot{Y}_2|\ddot{X}=x, \dot{Y}_1=y_1) = \\ &\quad\sum\nolimits_z \mathrm{E}(\dot{Y}_2|\ddot{X}=x, \dot{Z}=z, \dot{Y}_1=y_1)\,\Pr(\dot{Z}=z|\ddot{X}=x, \dot{Y}_1=y_1)\end{aligned} \tag{5.8}$$

Here the conditional distribution of $\dot{Z}$ that is used for mixing the conditional effects no longer only depends on exogenous variables ($\ddot{X}$ in this example), but also on the endogenous variable $\dot{Y}_1$. In general, this also changes the relationship between $\dot{Z}$ and $\ddot{X}$. For example, even if $\dot{Z}$ is independent of $\ddot{X}$, this is no longer true conditional on values of $\dot{Y}_1$. (This is illustrated by the data in Table 5.1.) Whether this provides a reason for thinking that the effect of $\ddot{X}$ in the reduced model is in some sense 'biased' will be discussed in **5.1.7**.

6. Example with linear regression functions

For further illustration of consequences of omitted variables, I still refer to Model 5.1 but assume linear functional relationships:

$$\mathrm{E}(\dot{Y}_1|x,z) = \alpha_1 + x\beta_1 + z\gamma_1 \tag{5.9}$$

$$\mathrm{E}(\dot{Y}_2|x,z,y_1) = \alpha_2 + x\beta_2 + z\gamma_2 + y_1\delta \tag{5.10}$$

As before, the interest concerns effects of x when z is omitted. In order to derive reduced models, I assume that the distribution of the values of z is given by a density function $f(z|x)$ with mean μ_z which, for simplicity, does not depend on x.

(a) I begin with the effect of x on the expectation of $\dot{Y}_1$. The relationship in the reduced model is

$$\mathrm{E}(\dot{Y}_1|x) = \int_z \mathrm{E}(\dot{Y}_1|x,z)\,f(z|x)\,dz$$

and one immediately finds

$$\mathrm{E}(\dot{Y}_1|x) = (\alpha_1 + \mu_z\gamma_1) + x\beta_1$$

showing that the parameter associated with x does not change. (This would not be true if (5.9) contained an interaction between x and z.)

(b) I now consider unconditional effects of x and z on $\dot{Y}_2$. The relationship in the reduced model is then given by

$$\mathrm{E}(\dot{Y}_2|x,z) = \int_{y_1} \mathrm{E}(\dot{Y}_2|x,z,y_1)\,f(y_1|x,z)\,dy_1$$

where $f(y_1|x,z)$ is the conditional density of $\dot{Y}_1$. Entailed by (5.9), this is a density with mean $\alpha_1 + x\beta_1 + z\gamma_1$. It follows that

$$\mathrm{E}(\dot{Y}_2|x,z) = (\alpha_2 + \alpha_1\delta) + x\,(\beta_2 + \beta_1\delta) + z\,(\gamma_2 + \gamma_1\delta) \tag{5.11}$$

The parameters are obviously different from the original parameters in (5.10). However, they are not 'biased', but correctly express the total (unconditional) effects of x and z on the expectation of $\dot{Y}_2$.

(c) I now consider the unconditional effect of x on $\dot{Y}_2$ when z is omitted. The relationship in the reduced model is then given by

$$\mathrm{E}(\dot{Y}_2|x) = \int_{y_1,z} \mathrm{E}(\dot{Y}_2|x,z,y_1)\,f(y_1,z|x)\,dy_1\,dz$$

where $f(y_1,z|x)$ is the joint density of y_1 and z. Since

$$f(y_1,z|x) = f(y_1|x,z)\,f(z|x)$$

one can continue with the derivation in (b):

$$\begin{aligned}\mathrm{E}(\dot{Y}_2|x) &= \int_z \mathrm{E}(\dot{Y}_2|x,z)\,f(z|x)\,dz \\ &= \alpha_2 + \alpha_1\delta + \mu_z\,(\gamma_2 + \gamma_1\delta) + x\,(\beta_2 + \beta_1\delta)\end{aligned} \tag{5.12}$$

The parameter associated with x is the same as in (5.11) where z was not omitted and has the same interpretation.

(d) Finally, I consider the conditional effect of x on $\dot{Y}_2$ when z is omitted. The relationship in the reduced model is then given by

$$\begin{aligned}\mathrm{E}(\dot{Y}_2|x,y_1) &= \int_z \mathrm{E}(\dot{Y}_2|x,z,y_1)\,f(z|x,y_1)\,dz \\ &= \alpha_2 + x\,\beta_2 + y_1\delta + \mathrm{E}(\dot{Z}|x,y_1)\,\gamma_2\end{aligned} \tag{5.13}$$

where $f(z|x,y_1)$ is the density of z conditional on x and y_1, and

$$\mathrm{E}(\dot{Z}|x,y_1) = \int_z z f(z|x,y_1)\,dz$$

is the conditional mean of z. How parameters change is best seen when using a linear relationship

$$\mathrm{E}(\dot{Z}|x,y_1) = \alpha_z + x\beta_z + y_1\delta_z \tag{5.14}$$

Inserting this into (5.13) leads to

$$\mathrm{E}(\dot{Y}_2|x,y_1) = (\alpha_2 + \alpha_z\gamma_2) + x\,(\beta_2 + \beta_z\gamma_2) + y_1\,(\delta + \delta_z\gamma_2) \qquad (5.15)$$

Parameters now consist of two parts. A first part equals the parameters as presupposed in (5.10), a second part consists of the contribution to estimating the conditional mean of the omitted variable, $\mathrm{E}(\dot{Z}|x,y_1)$.

To illustrate, assume $\alpha_1 = 1$, $\beta_1 = 1$, $\gamma_1 = 1$, $\alpha_2 = 0.5$, $\beta_2 = 1.5$, $\gamma_2 = 2$, $\delta = 0.2$, and $\mu_z = 1$. One then finds: $\alpha_z = 0$, $\beta_z = -0.5$, $\delta_z = 0.5$, entailing that the coefficient of x in (5.15) is less than β_2: $\beta_2 + \beta_z\gamma_2 = 1.5 - 0.5 \cdot 2$, and the coefficient of y_1 is greater than δ: $\delta + \delta_z\gamma_2 = 0.2 + 0.5 \cdot 2$.

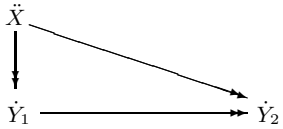
Of course, there is nothing wrong with these parameters. Assume that one knows an individual's values of $\ddot{X}$ and $\dot{Y}_1$, but not of $\ddot{Z}$. In order to apply the original model (5.10), one would need an estimate of the individual's value of $\ddot{Z}$. If one uses the conditional expectation as defined in (5.14), then (5.10) and (5.15) lead to the same result. In this sense, the difference in the parameters is required in order to get correct estimates with the reduced model.

7. Are effect definitions 'biased'?

When compared with (5.10), it is tempting to say that conditional effects calculated with (5.15) are 'biased'. However, one should be clear about what this possibly could mean.

(a) Not only the estimation, already the definition of an effect must be based on assuming a particular model. For example, one can use the following

Model 5.2



with the same variables as used for Model 5.1 to *define* an effect

$$\Delta^s(\dot{Y}_2; \ddot{X}[x', x''], \dot{Y}_1 = y) := \qquad (5.16)$$
$$\mathrm{E}(\dot{Y}_2|\,\ddot{X} = x'', \dot{Y}_1 = y) - \mathrm{E}(\dot{Y}_2|\,\ddot{X} = x', \dot{Y}_1 = y)$$

This effect compares the outcomes, in the second learning frame, $\dot{Y}_2$, of two *generic* individuals, one with $\ddot{X} = x'$ and the other one with $\ddot{X} = x''$. As made explicit in the definition, both have the same educational outcome, $\dot{Y}_1 = y$, in the first learning frame. However, no further assumptions are made about the individuals; they can differ in all variables not explicitly represented in the model (e.g. $\ddot{Z}$).[2]

(b) The statement that the effect defined in (5.16) is 'biased' has no meaning without explicitly referring to another model that can be used as a standard of comparison. For example, it might be said that the effect is biased because one *should* use Model 5.1 to define a 'less biased' effect. Such a normative statement is also required in order to call Model 5.2 a 'misspecified model'.[3]

(c) It is not possible to find, or generate, data for estimating Model 5.2 which entail that conditional distributions of the values of omitted variables (e.g. $\ddot{Z}$) are independent of $\ddot{X}$ and $\dot{Y}_1$. In order to assess the theoretically posited bias one would need to actually estimate the more comprehensive model 5.1, and this would require to get data about $\ddot{Z}$.

(d) Equation (5.13) seems to suggest that including estimates of the conditional mean values, $\mathrm{E}(\dot{Z}|x,y_1)$, into the reduced model would allow one to find an 'unbiased' estimate of β_2. This idea was proposed by Heckman (1979) who showed that in a class of specifically parameterized linear models one can find estimates of the conditional mean values of $\dot{Z}$ without observing its values. However, the required assumptions are very restrictive, and it has been shown that the approach can easily lead to very misleading parameter estimates (LaLonde 1986, Briggs 2004).

(e) It is a quite specific property of the simple linear model (5.10) that effects of $\ddot{X}$, $\ddot{Z}$ and $\dot{Y}_1$ are context-independent. In general, there are no context-independent effects to which effects derived from reduced models can be compared. In other words, then, already the presupposition that an effect can be associated uniquely with a single variable is misleading.

[2] This understanding of 'functional effects' is therefore quite different from effect definitions in the potential outcomes framework which are conceptually linked to particular individuals.

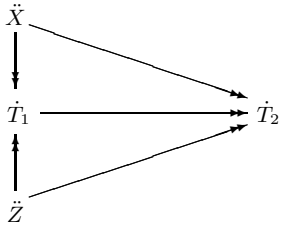
[3] This formulation is used, e.g., by Cameron and Heckman (1998).

5.2 Sequential transitions

1. Sequential binary choices

I now turn to models of sequential transitions. I begin with considering three consecutive learning frames, σ_j $(j = 1, 2, 3)$, and two binary variables, $\dot{T}_1$ and $\dot{T}_2$: $\dot{T}_j = 1$ if there is a transition from σ_j to σ_{j+1}, and $\dot{T}_j = 0$ if there is no transition into another learning frame. I consider the following

Model 5.3



The model has the same structure as Model 5.1, with $\dot{T}_j$ instead of $\dot{Y}_j$, but there is now the constraint

$$\Pr(\dot{T}_2{=}1 \,|\, \ddot{X}{=}x, \ddot{Z}{=}z, \dot{T}_1{=}0) = 0 \tag{5.17}$$

and from this follows

$$\begin{aligned}&\mathrm{E}(\dot{T}_2|\, \ddot{X}{=}x, \ddot{Z}{=}z) = \\ &\quad \mathrm{E}(\dot{T}_2|\, \ddot{X}{=}x, \ddot{Z}{=}z, \dot{T}_1{=}1)\, \mathrm{E}(\dot{T}_1|\, \ddot{X}{=}x, \ddot{Z}{=}z)\end{aligned} \tag{5.18}$$

2. Consideration of possible effects

In the same way as was discussed in the previous section, one can distinguish conditional and unconditional effects of the exogenous variables on $\dot{T}_2$. For example, the conditional effect of $\ddot{X}$ is

$$\begin{aligned}&\Delta^s_c(\dot{T}_2; \ddot{X}[x', x''], \ddot{Z}{=}z, \dot{T}_1{=}1) := \\ &\quad \mathrm{E}(\dot{T}_2|\, \ddot{X}{=}x'', \ddot{Z}{=}z, \dot{T}_1{=}1) - \mathrm{E}(\dot{T}_2|\, \ddot{X}{=}x', \ddot{Z}{=}z, \dot{T}_1{=}1)\end{aligned} \tag{5.19}$$

and the unconditional effect is

$$\begin{aligned}&\Delta^s_u(\dot{T}_2; \ddot{X}[x', x''], \ddot{Z}{=}z) := \\ &\quad \mathrm{E}(\dot{T}_2|\, \ddot{X}{=}x'', \ddot{Z}{=}z) - \mathrm{E}(\dot{T}_2|\, \ddot{X}{=}x', \ddot{Z}{=}z)\end{aligned} \tag{5.20}$$

In general, both conditional and unconditional effects are context-dependent, that is, depend on values of $\ddot{Z}$. To illustrate, I use the fictitious data in Table 5.2:

Table 5.2 Fictitious data for the sequential transition model.

x	z	t_1	t_2	cases	x	z	t_1	t_2	cases
0	0	0	0	500	1	0	0	0	300
0	0	0	1	0	1	0	0	1	0
0	0	1	0	200	1	0	1	0	140
0	0	1	1	300	1	0	1	1	560
0	1	0	0	400	1	1	0	0	100
0	1	0	1	0	1	1	0	1	0
0	1	1	0	200	1	1	1	0	90
0	1	1	1	400	1	1	1	1	810

Effects on first transition:

$$\Delta^s(\dot{T}_1; \ddot{X}[0,1], \ddot{Z}{=}0) = 0.70 - 0.50 = 0.20$$
$$\Delta^s(\dot{T}_1; \ddot{X}[0,1], \ddot{Z}{=}1) = 0.90 - 0.60 = 0.30$$

Conditionals effect on second transition:

$$\Delta^s_c(\dot{T}_2; \ddot{X}[0,1], \ddot{Z}{=}0, \dot{T}_1{=}1) = 0.80 - 0.60 = 0.20$$
$$\Delta^s_c(\dot{T}_2; \ddot{X}[0,1], \ddot{Z}{=}1, \dot{T}_1{=}1) = 0.90 - 0.67 = 0.23$$

Unconditional effect on second transition:

$$\Delta^s_u(\dot{T}_2; \ddot{X}[0,1], \ddot{Z}{=}0) = 0.56 - 0.30 = 0.26$$
$$\Delta^s_u(\dot{T}_2; \ddot{X}[0,1], \ddot{Z}{=}1) = 0.81 - 0.40 = 0.41$$

How to think of possible effects of $\dot{T}_1$? These effects are already defined by the set-up of the model and need not to be estimated. $\dot{T}_1 = 0$ deterministically entails $\dot{T}_2 = 0$. $\dot{T}_1 = 1$, on the other hand, is a necessary condition for $\dot{T}_2 = 1$, but has no further causal effect on the probability distribution of $\dot{T}_2$. The role played by $\dot{T}_1$ in Model 5.3 is therefore quite different from the role played by $\dot{Y}_1$ in Model 5.1.

3. Models in terms of latent variables

Parametric models for the transition variables, $\dot{T}_j$, are often set up in terms of latent variables. This allows formulations which, in a sense, parallel (5.9) and (5.10), most easily in pseudo-indeterministic forms. For example:

$$\dot{T}^*_1 = \alpha_1 + x\beta_1 + z\gamma_1 + \dot{U}_1 \tag{5.21}$$
$$\dot{T}^*_2 = \alpha_2 + x\beta_2 + z\gamma_2 + \dot{U}_2 \tag{5.22}$$

$\dot{T}^*_1$ and $\dot{T}^*_2$ are the latent variables which are linked to the binary transition variables by $\dot{T}_1 = I[\dot{T}^*_1 \geq 0]$ and $\dot{T}_2 = I[\dot{T}^*_2 \geq 0]$. $\dot{U}_1$ and $\dot{U}_2$ are random variables with distribution functions F_1 and F_2, respectively. The distributions are most often assumed to be symmetrical around a zero mean. This

then entails:

$$\mathrm{E}(\dot{T}_1|x,z) = F_1(\alpha_1 + x\beta_1 + z\gamma_1) \tag{5.23}$$

$$\mathrm{E}(\dot{T}_2|x,z,\dot{T}_1{=}1) = F_2(\alpha_2 + x\beta_2 + z\gamma_2) \tag{5.24}$$

showing that F_1 and F_2 simply serve as functions linking the explanatory variables to the binomial distributions of the outcome variables (see also Section 3.1).

However, even when using latent variables there remains an essential difference between models 5.1 and 5.3. Model 5.1 relates to a generic individual that participates in both learning frames, and this allows one to think of a joint distribution of the two educational outcomes, $\dot{Y}_1$ and $\dot{Y}_2$. Now consider the sequential transition model 5.3. In this model, $\dot{T}_1 = 0$ deterministically entails $\dot{T}_2{=}0$. While one can nevertheless think of a joint distribution of $\dot{T}_1$ and $\dot{T}_2$, there is no corresponding joint distribution of the latent variables, $\dot{T}_1^*$ and $\dot{T}_2^*$. If $\dot{T}_1^* < 0$, there is no transition and consequently no second latent variable, $\dot{T}_2^*$.

The argument shows that model specifications for the sequential transition model cannot sensibly begin with assuming a joint distribution for latent transition variables.[4]

4. Describing selection processes

In the sequential transition model, the transition variables can be interpreted as representing selections: if $\dot{T}_j = 1$, an individual is selected into a subsequent learning frame. However, in order to actually consider a selection process one needs a collection of individuals, say Ω, and statistical variables corresponding to the model's exogenous variables. As an example, one can use the data in Table 5.2 which provide values for

$$(X, Z, T_1, T_2) : \Omega \longrightarrow \{0,1\}^4 \tag{5.25}$$

Based on information about these variables one can describe how the distributions of the exogenous variables change in the consecutive learning frames. In this example:

	t_1	t_2	$\mathrm{P}(X{=}1\|T_1{=}t_1, T_2{=}t_2)$	$\mathrm{P}(Z{=}1\|T_1{=}t_1, T_2{=}t_2)$
σ_1	0	0	0.308	0.385
σ_2	1	0	0.365	0.460
σ_3	1	1	0.662	0.585

Assuming that $X = 1$ represents 'high SES', and $Z = 1$ 'high aspiration', this would mean that the selection process leads to larger proportions of individuals with 'high SES' and 'high aspiration'.

In order to formulate a general relationship between effects on transitions and subsequent selections, I consider the unconditional effects

$$\Delta^s_u(\dot{T}_j; \ddot{X}[x', x''], \ddot{Z}{=}z) := \mathrm{E}(\dot{T}_j|\,\ddot{X}{=}x'', \ddot{Z}{=}z) - \mathrm{E}(\dot{T}_2|\,\ddot{X}{=}x', \ddot{Z}{=}z) \tag{5.26}$$

Given corresponding statistical variables, as defined in (5.25), there is the following simple relationship:

$$\Delta^s_u(\dot{T}_j; \ddot{X}[x', x''], \ddot{Z}{=}z) > 0 \Longleftrightarrow \frac{\mathrm{P}(X{=}x''|T_j{=}1, Z{=}z)}{\mathrm{P}(X{=}x'|T_j{=}1, Z{=}z)} > \frac{\mathrm{P}(X{=}x''|Z{=}z)}{\mathrm{P}(X{=}x'|Z{=}z)}$$

This is easily understandable: If individuals with $\ddot{X} = x''$ have a higher probability for making the transition $T_j = 1$ than individuals with $\ddot{X} = 0$, they will become relatively more frequent conditional on $T_j{=}1$.

Note, however, that this relationship has no implications for a comparison of effects *across* transitions. In general, the conditional effect defined in (5.19) can be greater or smaller than the effect of $\ddot{X}$ on the first transition.

5. Consequences of omitted variables

Also in the sequential transition model one can think about omitted variables. This is often a concern in the literature.[5] Considerations parallel the discussion of omitted variables in the models of sequential educational outcomes, and the results of **5.1.5** and the conclusions of **5.1.7** can be applied without any essential modification.

It might nevertheless be interesting to consider a further example. For this example, I start from (5.23) and (5.24) and assume logit specifications. Corresponding to $\ddot{X}$ I assume a binary variable, X, with $\mathrm{P}(X{=}1) = 0.5$; and corresponding to $\ddot{Z}$ I assume a continuous variable, Z, with a standard normal density, $f(z)$, independent of X. For the parameters I choose $\alpha_1 = \beta_1 = \gamma_1 = 1$ and $\alpha_2 = \beta_2 = \gamma_2 = 1$.

Conditional on $T_1{=}1$ both distributions change. Instead of $\mathrm{P}(X{=}1) = 0.5$, there is a higher value $\mathrm{P}(X{=}1|T_1{=}1) = 0.55$. How the distribution of Z changes is shown in Figure 5.1. The figure also shows that X and Z are no longer independent. Both conditional mean values have increased: $\mathrm{M}(Z|T_1 = 1, X = 0) = 0.255$, and $\mathrm{M}(Z|T_1 = 1, X = 1) = 0.137$. The difference is due to a negative correlation between X and Z which, in this example, results from conditioning on T_1.

[4] As an example, I refer to Holm and Jaeger (2011) who proposed to begin with assuming a joint normal distribution for the latent transition variables.

[5] E.g., Mare (1993), Cameron and Heckman (1998), Holm and Jaeger (2011).

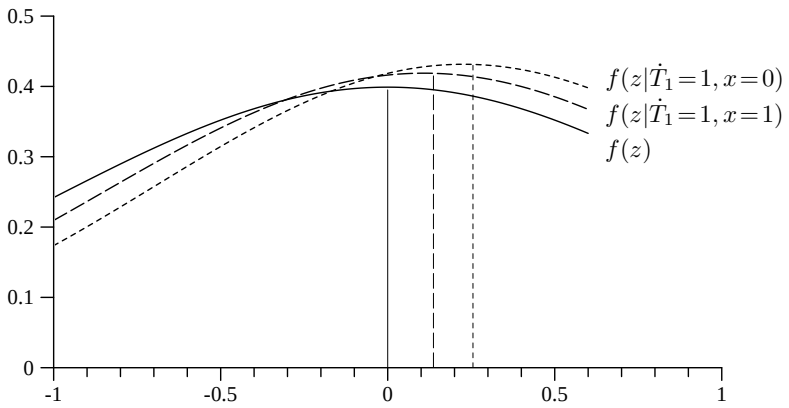


Figure 5.1 Unconditional and conditional density functions of Z and their mean values as described in the text.

Now assume that Z is not observed. The conditional effect of $\ddot{X}$ on the second transition that can be estimated is then given by

$$\mathrm{E}(\dot{T}_2|\ddot{X}=1,\dot{T}_1=1)-\mathrm{E}(\dot{T}_2|\ddot{X}=0,\dot{T}_1=1)= \tag{5.27}$$
$$\int_z \mathrm{E}(\dot{T}_2|\ddot{X}=1,\dot{Z}=z,\dot{T}_1=1)\,f(z|\ddot{X}=1,\dot{T}_1=1)\,dz-$$
$$\int_z \mathrm{E}(\dot{T}_2|\ddot{X}=0,\dot{Z}=z,\dot{T}_1=1)\,f(z|\ddot{X}=0,\dot{T}_1=1)\,dz$$

In our example, this effect has the value

$$0.1185 = 0.8630 - 0.7445$$

Of course, one can argue that this effect is not only due to $\ddot{X}$ but also to the selection effect that shows up in the different conditional distributions of $\dot{Z}$. In our example, these effects can be separated by using $f(z|x)$ instead of $f(z|x,\dot{T}_1=1)$. Instead of (5.27), one then calculates a counterfactual effect with

$$\int_z \mathrm{E}(\dot{T}_2|\ddot{X}=1,\dot{Z}=z,\dot{T}_1=1)\,f(z|\ddot{X}=1)\,dz- \tag{5.28}$$
$$\int_z \mathrm{E}(\dot{T}_2|\ddot{X}=0,\dot{Z}=z,\dot{T}_1=1)\,f(z|\ddot{X}=0)\,dz$$

In our example, one finds the value

$$0.1478 = 0.8445 - 0.6967$$

which is obviously larger. The difference can be attributed to the selection effect.

However, it would not be sensible to think of the effect calculated with (5.27) as a 'biased estimate' of the effect defined by (5.28). The latter effect is purely counterfactual, and could only be calculated if $\ddot{Z}$ had been observed.

Chapter 6

Educational outcomes and transitions

The models discussed in Section 5.2 make transition probabilities directly dependent on variables representing the family background (e.g. parents' SES and educational levels). It is highly plausible, however, that transition probabilities also depend on the educational outcomes in earlier learning frames. This idea has lead to a longstanding debate about how effects of the familiy background are mediated.

Following Boudon (1974), researchers often focus on a distinction between primary and secondary effects (see, e.g., Baumert et al. 2003, Erikson et al. 2005, Jackson et al. 2007, Erikson and Rudolphi 2010, Schindler and Reimer 2010, Neugebauer 2010). I take up this discussion in the first section and consider two different modeling approaches.

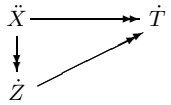
In the second section, I discuss an enlarged version of the sequential transition model of Section 5.2 in which a distinction between primary and secondary effects is explicitly represented.

6.1 Primary and secondary effects

1. An often used modeling approach

Researchers often use a model that basically has the following structure:

Model 6.1.1



$\dot{T}$ is a binary dependent variable representing the transition of interest, for example, $\dot{T}=1$ if there is a transition from lower to higher secondary education, and $\dot{T}=0$ otherwise. There are two main explanatory variables:

- $\ddot{X}$ is an exogenous explanatory variable that represents some aspect of the family background, e.g., fathers occupational class (Erikson et al. 2005), or parents' educational level (Kloosterman et al. 2009, Neugebauer 2010).
- $\dot{Z}$ is an endogenous explanatory variables which represents the 'level of academic performance' an individual has reached in the situation where the transition is to be made. Values are often based on a scholastic aptitude test.

This model is then used to distinguish two kinds of effects of $\ddot{X}$. A *primary effect* concerns the dependence of $\dot{Z}$ on $\ddot{X}$. For example, comparing x' and x'', the primary effect could be defined as

$$\mathrm{E}(\dot{Z}|\ddot{X}=x'') - \mathrm{E}(\dot{Z}|\ddot{X}=x')$$

The *secondary effect* concerns the effect of $\ddot{X}$ on $\dot{T}$, while $\dot{Z}$ is in some way fixed. For example,

$$\Pr(\dot{T}=1|\ddot{X}=x'', \dot{Z}=z) - \Pr(\dot{T}=1|\ddot{X}=x', \dot{Z}=z)$$

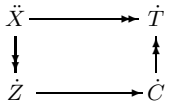
2. Representing transition opportunities

Model 6.1.1 presupposes a direct link between 'academic performance' and the transition variable. A different modeling approach attempts to represent the individual's choice situation. I assume that this can be done with a binary variable, say $\dot{C}$, such that $\dot{C}=1$ if an individual has the choice to opt for the transition of interest. So one can distinguish between

- the dependence of $\dot{C}$ on $\ddot{X}$ ('primary effect'), and
- the dependence of $\dot{T}$ on $\ddot{X}$ given that $\dot{C}=1$ ('secondary effect').

Assuming that $\dot{C}$ depends in some way on 'academic performance', $\dot{Z}$, leads one to the following model:

Model 6.1.2



Instead of a direct link between $\dot{Z}$ and $\dot{T}$, there is now a mediating binary variable, $\dot{C}$. For subsequent illustrations, I assume a simple deterministic relationship between $\dot{Z}$ and $\dot{C}$:

$$\dot{C} = I[\dot{Z} \geq z_0]$$

where z_0 is a threshold value. $I[\ldots]$ denotes the indicator function; in this example: $I[\dot{Z} \geq z_0]$ takes the value 1 if $\dot{Z} \geq z_0$ and is otherwise zero.

Instead of the definitions given in **6.1.1**, one now gets quite different expressions for primary and secondary effects. Assuming that $\dot{Z}$ is a discrete variable, the primary effect is given by

$$\Pr(\dot{C}{=}1|\ddot{X}{=}x'') - \Pr(\dot{C}{=}1|\ddot{X}{=}x') = \sum\nolimits_{z \geq z_0} \big(\Pr(\dot{Z}{=}z|\ddot{X}{=}x'') - \Pr(\dot{Z}{=}z|\ddot{X}{=}x')\big)$$

And, since the model entails that $\Pr(\dot{T}{=}1|\dot{C}{=}0) = 0$, the secondary effect is given by

$$\Pr(\dot{T}{=}1|\ddot{X}{=}x'', \dot{C}{=}1) - \Pr(\dot{T}{=}1|\ddot{X}{=}x', \dot{C}{=}1)$$

Notice that, given $\dot{C}$, this effect is now independent of $\dot{Z}$. (A model that assumes an additional dependence of $\dot{T}$ on $\dot{Z}$ will be considered below.)

3. Numerical illustration with simulated data

The two models can lead to quite different conclusions. For illustration, I use Model 6.1.2 to create artificial data. $\ddot{X} \in \{0, 1, 2, 3, 4\}$. For $\ddot{X} = x$, $\dot{Z}$ is normally distributed with mean μ_x and variance 1. Threshold value: $z_0 = -1$. Table 6.1.1 shows the values assumed for μ_x in column 2, and the assumed conditional probabilities $\Pr(\dot{T}{=}1 \,|\, \dot{C}{=}1, \ddot{X}{=}x)$ in column 4 (these are the values resulting from the simulation). This allows deriving values for the remaining conditional probabilities. The relationship is given by

$$\Pr(\dot{T}{=}1|\ddot{X}{=}x) = \Pr(\dot{T}{=}1|\dot{C}{=}1, \ddot{X}{=}x)\,\Pr(\dot{C}{=}1|\ddot{X}{=}x)$$

(see the formally corresponding relationship in **1.3.2**). Obviously, in this example almost all effects are secondary.

Table 6.1.1 Values of conditional probabilities assumed in Model 6.1.2.

x	μ_x	$\Pr(\dot{C}{=}1 \,|\, \ddot{X}{=}x)$	$\Pr(\dot{T}{=}1 \,|\, \dot{C}{=}1, \ddot{X}{=}x)$	$\Pr(\dot{T}{=}1 \,|\, \ddot{X}{=}x)$
0	0.00	0.84	0.5	0.42
1	0.05	0.85	0.6	0.52
2	0.10	0.87	0.7	0.60
3	0.15	0.88	0.8	0.70
4	0.20	0.89	0.9	0.80

Now I assume Model 6.1.1 instead of 6.1.2 and use a logit model to specify a continuous dependence of the probability of $\dot{T}{=}1$ on values of $\dot{Z}$ and dummy variables $X_j := I[\ddot{X}{=}j]$:

$$\Pr(\dot{T}{=}1|\ddot{X}{=}x, \dot{Z}{=}z)) \approx \frac{\exp(X_0\beta_0 + X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + X_4\beta_4 + Z\gamma)}{1 + \exp(X_0\beta_0 + X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + X_4\beta_4 + Z\gamma)}$$

In order to calculate parameters, I use, for each value of $\ddot{X}$, 10000 cases, randomly generated according to the assumptions in Table 6.1.1.[1] The estimated parameters are:

$$\hat{\beta}_0 = -0.3696,\ \hat{\beta}_1 = 0.0503,\ \hat{\beta}_2 = 0.3719,\ \hat{\beta}_3 = 0.8352,\ \hat{\beta}_4 = 1.3610$$
$$\hat{\gamma}\ = 0.7586$$

Figure 6.1.1 shows, for $\ddot{X} = 1$ and $\ddot{X} = 4$, how the estimated probabilities depend on values of $\dot{Z}$.

Compared with Model 6.1.2, the results are quite different. In Model 6.1.2, if $\dot{Z} \geq z_0$, there is no further effect of $\dot{Z}$ on $\dot{T}$. The logit model, in contrast, suggests that $\dot{Z}$ has always an important effect, even conditional on values of $\ddot{X}$. For example:

$$\Pr(\dot{T}{=}1|\dot{Z}{=}1, \ddot{X}{=}1) - \Pr(\dot{T}{=}1|\dot{Z}{=}0, \ddot{X}{=}1) = 0.2$$
$$\Pr(\dot{T}{=}1|\dot{Z}{=}1, \ddot{X}{=}4) - \Pr(\dot{T}{=}1|\dot{Z}{=}0, \ddot{X}{=}4) = 0.1$$

Correspondingly, the logit model underestimates secondary effects. Consider, for example, the effect

$$\Pr(\dot{T}{=}1|\dot{Z}{=}1, \ddot{X}{=}4) - \Pr(\dot{T}{=}1|\dot{Z}{=}1, \ddot{X}{=}1)$$

While its value is $0.9 - 0.6 = 0.3$ in Model 6.1.2, the logit model suggests the value 0.2.

[1] Individual values of $\dot{T}$, say t_i for an individual i with covariate values x_i and z_i, are generated as follows: $t_i = 1$ if $z_i \geq -1$ and $r_i \leq \Pr(\dot{T}{=}1 \,|\, \dot{C}{=}1, \ddot{X}{=}x_i)$, where r_i is a random number equally distributed in $[0, 1]$; otherweise $t_i = 0$.

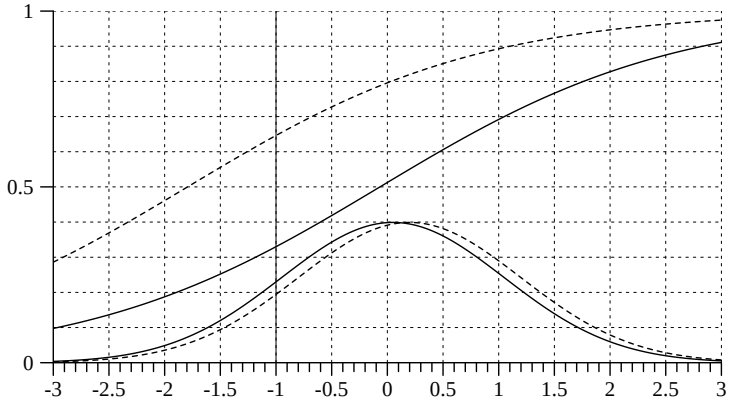


Figure 6.1.1 Distribution of $\dot{Z}$ for $\ddot{X}=1$ (solid line) and $\ddot{X}=4$ (dashed line). Also shown: $\exp(\hat{\beta}_1 + z\hat{\gamma})/(1 + \exp(\hat{\beta}_1 + z\hat{\gamma}))$ (solid line) and $\exp(\hat{\beta}_4 + z\hat{\gamma})/(1 + \exp(\hat{\beta}_4 + z\hat{\gamma}))$ (dashed line).

4. Hypothetical effect combinations

In order to illustrate the relative importance of primary and secondary effects, some authors have suggested to consider hypothetical effect combinations (Erikson et al. 2005, Jackson et al. 2007, Kloosterman et al. 2009, Schindler and Reimer 2010). This can easily be done with Model 6.1.2. One can calculate

$$H'(x,x') := \Pr(\dot{T}=1 \,|\, \dot{C}=1, \ddot{X}=x') \Pr(\dot{C}=1 \,|\, \ddot{X}=x)$$

and interpret this as the probability of $\dot{T}=1$ under the assumption that the primary effect is given by $\ddot{X}=x$ and the secondary effect is given by $\ddot{X}=x'$.

A similar calculation can be done with Model 6.1.1 by exploiting the knowledge of the conditional distributions of $\dot{Z}$. In our example, it was assumed that, depending on $\ddot{X}=x$, $\dot{Z}$ is normally distributed with density $\phi(z;\mu_x,1)$. So one can calculate

$$H(x,x') := \int_z \Pr(\dot{T}=1 \,|\, \ddot{X}=x', \dot{Z}=z)\, \phi(z;\mu_x,1)\, dz \qquad (6.1)$$

and interpret this as the probability of $\dot{T}=1$ under the assumption that the primary effect is given by x and the secondary effect is given by x'.

The quantities $H(x,x')$ and $H'(x,x')$ can be used for a decomposition of the total effect which is analogous to the decomposition into a direct and

Table 6.1.2 Hypothetical probabilities of $\dot{T}=1$ when assuming that primary effects depend on $\ddot{X}=x$ and secondary effects depend on $\ddot{X}=x'$. Calculations based on data in Table 6.1.1.

	Model 6.1.1		Model 6.1.2	
x	$H(x,1)$	$H(x,4)$	$H'(x,1)$	$H'(x,4)$
0	0.513	0.773	0.425	0.756
1	0.521	0.779	0.510	0.765
2	0.529	0.785	0.595	0.783
3	0.536	0.789	0.680	0.792
4	0.545	0.795	0.765	0.801

an indirect effect that was discussed in **4.2.6**:

$$H(x'',x'') - H(x',x') = \qquad (6.2)$$
$$[H(x'',x'') - H(x',x'')] + [H(x',x'') - H(x',x')]$$

(analoguous for H'). The first term on the right-hand side is the primary effect that corresponds to the indirect effect of $\ddot{X}$, the second term is the secondary effect that corresponds to the direct effect of $\ddot{X}$.

Values of $H(x,x')$ and $H'(x,x')$ for our example are shown in Table 6.1.2. To illustrate, I consider decompositions of $H(4,4) - H(1,1)$:

	primary	secondary	total
Model 6.1.1	0.016	0.258	0.274
Model 6.1.2	0.036	0.255	0.291

This confirms that in our example the secondary effect is much greater than the primary effect.

5. Comparing the two models

Models 6.1.1 and 6.1.2 are different in several respects. First, there are different data requirements. Model 6.1.1 only requires some measure of 'academic performance'. This can be based on institutionalized scholastic aptitude tests, or tests developed and performed by a researcher. In contrast, the primarily important data requirement for Model 6.1.2 concerns the variable $\dot{C}$ which represents the choice situation.

A second point concerns the interpretation. Model 6.1.2 is based on a conceptually clear distinction between primary and secondary effects: primary effects concern the generation of opportunities, secondary effects concern the actually performed choices. There is no correspondingly clear

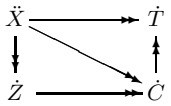
distinction in Model 6.1.1. The idea that the chosen transition in some way depends on 'academic performance' is not easily interpretable.

This reflects a deeper distinction between the two models. Model 6.1.2 can be understood as an analytical model that aims to understand effects of the family background both on the generation and use of educational transition opportunities. In contrast, Model 6.1.1 aims to attribute the actually performed transitions to two sources: a 'meritocratic' source represented by 'academic performance', and any further factors which cannot be justified from a meritocratic view.[2]

6. Model specification and observed correlations

Model 6.1.2 assumes that $\dot{C}$, representing the choice situation, solely depends on 'academic performance' ($\dot{Z}$). It is well possible, of course, that observations suggest that $\dot{C}$ depends on $\ddot{X}$ in ways not mediated through $\dot{Z}$ (see Ditton et al. (2005) for some empirical evidence). One might then consider the following modified model:

Model 6.1.3

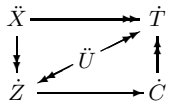


There is then no longer a deterministic relationship between $\dot{C}$ and $\dot{Z}$; and the 'primary effect' of $\ddot{X}$ (as defined at the beginning of this section) cannot be interpreted solely in terms of 'academic performance'. Nevertheless, it is still possible to separate a primary and a secondary effect.

But now assume that observations suggest a correlation between $\dot{Z}$ and $\dot{T}$ even conditional on $\dot{C}=1$. There are then different possibilities to modify Model 6.1.2, corresponding to different theoretical ideas. One possibility is to add an arrow from $\dot{Z}$ to $\dot{T}$. This would reflect the hypothesis that the degree of 'academic performance' not only determines the possibility of a choice, but in addition influences the final decision. A quite different possibility is to suppose an unobserved variable, say $\ddot{U}$, which, like $\ddot{X}$ influences both $\dot{Z}$ and $\dot{T}$. This would suggest the following model:

[2]Sometimes this interest in meritocratic evaluations is explicitly mentioned, see e.g. Schindler and Reimer (2010: 624).

Model 6.1.4



The example demonstrates that it is insufficient to think in terms of correlation (observed or theoretically postulated). There are always several different dependency relations between variables that can generate the observed correlations.

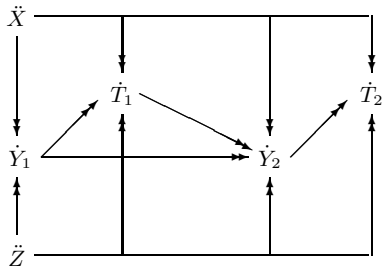
6.2 An enlarged transition model

I now consider an enlarged version of the transition model that was discussed in Section 5.2 which includes effects of the educational outcome of a foregoing learning frame.

1. Including educational outcomes

I consider the following model which is a combination of the models considered in sections 5.1 and 5.2:

Model 6.2.1



As before, $\dot{Y}_j$ is the educational outcome, and $\dot{T}_j$ represents the transition. The model entails the following constraint: $\dot{Y}_2$ and $\dot{T}_2$ are only defined if $\dot{T}_1 = 1$. This must be taken into account when defining conditional effects on $\dot{Y}_2$ and $\dot{T}_2$.

2. Specification of dependency relations

Remembering the discussion of primary and secondary effects in Section 6.1, a critical question concerns how to specify the relationship between the educational outcome of a learning frame, $\dot{Y}_j$, and the subsequent transition,

$\dot{T}_j$. Here I assume that a minimal level, say λ_j, is a necessary condition for making a transition:

$$\dot{T}_j = 0 \text{ if } \dot{Y}_j < \lambda_j \tag{6.3}$$

and that the probability of $\dot{T}_j = 1$ depends in some way on the difference, $\dot{Y}_j - \lambda_j$. To illustrate a possible specification, I use

$$\begin{aligned}&\Pr(\dot{T}_j = 1 | \ddot{X} = x, \ddot{Z} = z, \dot{Y}_j = y_j) = \\ &\quad F_j(\alpha_j + x\beta_j + z\gamma_j + \eta_j(y_j - \lambda_j))\, I[y_j \geq \lambda_j]\end{aligned} \tag{6.4}$$

where F_j is a link function (e.g., a normal or logistic distribution function).

A further question concerns the specification of how educational outcomes depend on previous conditions. In contrast to the models considered in Section 5.1, it is no longer sufficient to focus only on the expectation of $\dot{Y}_j$; one needs at least a specification of how the probability of $\dot{Y}_j \geq \lambda_j$ depends on explanatory variables.

3. Primary and secondary effects

Model 6.2.1 can be used to consider the distinction between primary and secondary effects of SES that was introduced in Section 6.1. Primary effects concern the effect of $\ddot{X}$ on the opportunity to make a transition. For the first transition, this effect can be defined by

$$\begin{aligned}&\Delta^s(I[\dot{Y}_1 \geq \lambda_1]; \ddot{X}[x', x''], \ddot{Z} = z) := \\ &\quad \Pr(\dot{Y}_1 \geq \lambda_1 | \ddot{X} = x'', \ddot{Z} = z) - \Pr(\dot{Y}_1 \geq \lambda_1 | \ddot{X} = x', \ddot{Z} = z)\end{aligned} \tag{6.5}$$

Secondary effects, on the other hand, concern the effect of $\ddot{X}$ on actually making the transition, given that there is an opportunity. For the first transition, this effect can be defined by

$$\begin{aligned}&\Delta^s_c(\dot{T}_1; \ddot{X}[x', x''], \ddot{Z} = z, \dot{Y}_1 \geq \lambda_1) := \\ &\quad \Pr(\dot{T}_1 = 1 | \ddot{X} = x'', \ddot{Z} = z, \dot{Y}_1 \geq \lambda_1) - \Pr(\dot{T}_1 = 1 | \ddot{X} = x', \ddot{Z} = z, \dot{Y}_1 \geq \lambda_1)\end{aligned} \tag{6.6}$$

The secondary effect is conditional on $\dot{Y}_1 \geq \lambda_1$ and must be distinguished from the unconditional effect

$$\begin{aligned}&\Delta^s_u(\dot{T}_1; \ddot{X}[x', x''], \ddot{Z} = z) := \\ &\quad \Pr(\dot{T}_1 = 1 | \ddot{X} = x'', \ddot{Z} = z) - \Pr(\dot{T}_1 = 1 | \ddot{X} = x', \ddot{Z} = z)\end{aligned} \tag{6.7}$$

This is the effect that would be estimated when using the sequential transition model that was considered in Section 5.2 and which does not allow distinguishing between primary and secondary effects.

To illustrate, I assume that the relationship between $\dot{Y}_1$ and $\dot{T}_1$ is given by (6.4), and $\dot{Y}_1$ is normally distributed with mean

$$\mathrm{E}(\dot{Y}_1 | \ddot{X} = x, \ddot{Z} = z) = \alpha_1^y + x\beta_1^y + z\gamma_1^y \tag{6.8}$$

and a unit variance. The primary effect can then be calculated from

$$\Pr(\dot{Y}_1 \geq \lambda_1 | \ddot{X} = x, \ddot{Z} = z) = \Phi(\alpha_1^y + x\beta_1^y + z\gamma_1^y - \lambda_1) \tag{6.9}$$

(where Φ denotes the standard normal distribution function), the unconditional effect can be calculated from

$$\begin{aligned}&\Pr(\dot{T}_1 = 1 | \ddot{X} = x, \ddot{Z} = z) = \\ &\quad \int_{y \geq \lambda_1} \Pr(\dot{T}_1 = 1 | \ddot{X} = x, \ddot{Z} = z, \dot{Y}_1 = y)\, f(y|x, z)\, dy = \\ &\quad \int_{y \geq \lambda_1} F_1(\alpha_j + x\beta_j + z\gamma_j + \eta_j(y - \lambda_j))\, \phi(y - (\alpha_1^y + x\beta_1^y + z\gamma_1^y))\, dy\end{aligned}$$

(where ϕ denotes the standard normal density function), and the secondary effect can be calculated from

$$\Pr(\dot{T}_1 = 1 | \ddot{X} = x'', \ddot{Z} = z, \dot{Y}_1 \geq \lambda_1) = \frac{\Pr(\dot{T}_1 = 1 | \ddot{X} = x, \ddot{Z} = z)}{\Pr(\dot{Y}_1 \geq \lambda_1 | \ddot{X} = x, \ddot{Z} = z)}$$

For a numerical illustration, I use a logit specification for F_1 and assume $\alpha_1 = \beta_1 = \gamma_1 = \eta_1 = 1$, and $\alpha_1^y = 0.5$, $\beta_1^y = \gamma_1^Y = 0.7$, $\lambda_1 = 1$. For $z = 0$, one finds:

x	$\Pr(\dot{Y}_1 \geq \lambda_1 \mid x, z)$	$\Pr(\dot{T}_1 = 1 \mid x, z, \dot{Y}_1 \geq \lambda_1)$	$\Pr(\dot{T}_1 = 1 \mid x, z)$
-1	0.115	0.617	0.071
0	0.309	0.825	0.255
1	0.579	0.940	0.544
2	0.816	0.982	0.801

Effects depend on x, for example:

$$\begin{aligned}&\Delta^s(I[\dot{Y}_1 \geq \lambda_1]; \ddot{X}[0, 1], \ddot{Z} = 0) = 0.579 - 0.309 = 0.270 \\ &\Delta^s_c(\dot{T}_1; \ddot{X}[0, 1], \ddot{Z} = 0, \dot{Y}_1 \geq 1) \;= 0.940 - 0.825 = 0.115 \\ &\Delta^s_u(\dot{T}_1; \ddot{X}[0, 1], \ddot{Z} = 0) \qquad\quad = 0.544 - 0.255 = 0.289\end{aligned}$$

4. Educational outcomes and primary effects

Given the understanding that primary effects concern opportunities for transitions, they must be distinguished from effects on educational outcomes. This can be illustrated with the specifications assumed in the previous subsection. Using (6.8) for the educational outcome, $\dot{Y}_1$, the marginal effect of $\ddot{X}$ is β_1^y. In contrast, using (6.9), the marginal effect of $\ddot{X}$ on $I[\dot{Y}_1 \geq \lambda_1]$ is

$$\frac{\partial \Pr(\dot{Y}_1 \geq \lambda_1 | \ddot{X} = x, \ddot{Z} = z)}{\partial x} = \phi(\alpha_1^y + x\beta_1^y + z\gamma_1^y - \lambda_1)\, \beta_1^y \tag{6.10}$$

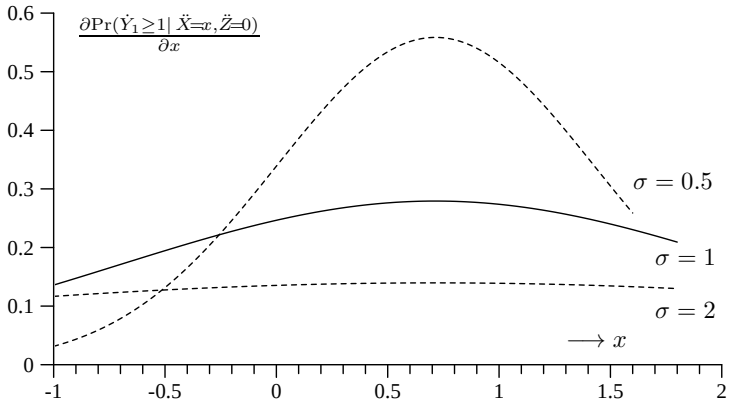


Figure 6.2.1 Dependence of the marginal effect (6.10), with $\ddot{Z} = 0$, on values of σ. Parameters: $\alpha_1^y = 0.5$, $\beta_1^y = \gamma_1^y = 0.7$, $\lambda_1 = 1$.

This shows that the primary effect is not only determined by β_1^y, but also by values of explanatory variables and, most important, by λ_1.

This also entails that the primary effect depends on the degree of inequality in the educational outcomes. This can be illustrated by introducing a parameter for the variance of $\dot{Y}_1$:

$$\dot{Y}_1 = \alpha_1^y + x\beta_1^y + z\gamma_1^y + \sigma\dot{U}_1 \tag{6.11}$$

where $\dot{U}_1$ is a normally distributed random variable with unit variance. The derivation of (6.9) was based on $\sigma = 1$. Starting from (6.11), one gets

$$\Pr(\dot{Y}_1 \geq \lambda_1 | \, \ddot{X}=x, \ddot{Z}=z) = \Phi\big(\frac{1}{\sigma}[\alpha_1^y + x\beta_1^y + z\gamma_1^y - \lambda_1]\big) \tag{6.12}$$

Consequences of an increasing variance of the educational outcomes depend on the expectation of $\dot{Y}_1$:

– if $\mathrm{E}(\dot{Y}_1|x,z) > \lambda_1$, the probability of $\dot{Y}_1 \geq \lambda_1$ decreases, and

– if $\mathrm{E}(\dot{Y}_1|x,z) < \lambda_1$, the probability of $\dot{Y}_1 \geq \lambda_1$ increases.

Also the marginal effects of the explanatory variables depend on the variance of the educational outcomes. Figure 6.2.1 illustrates how the marginal effect of $\ddot{X}$, as defined in (6.10), depends on values of σ. Notice that these effects also depend on values of $\ddot{Z}$.

Chapter 7

Selection between several alternatives

The models discussed in the two preceding chapters used binary outcome variables. I now consider situations where a selection between more than two alternatives can take place. The first section discusses how such situations can be modeled with multinomial logit models, and illustrates this with artificial data. This section also shows how one can take into account constraints on available alternatives. In the second section, I use data from the the AIDA survey (*Aufwachsen in Deutschland: Alltagswelten*), conducted by the *Deutsches Jugendinstitut* (Munich), to illustrate a multinomial logit model of the selection of school types in grade 5.

7.1 Models with several alternatives

1. The functional framework

The functional framework is simple:

$$(\ddot{X}, \ddot{Z}) \longrightarrow\!\!\!\!\!\rightarrow \dot{Y} \tag{7.1}$$

The dependent variable, $\dot{Y}$, can take values in a set $\mathcal{Y} = \{1, \ldots, m\}$, representing m alternatives. Many different interpretations are possible. Here I consider a situation where the alternatives represent learning frames, and the model concerns the selection of students into one of these learning frames. As an example, I consider selection into school types in grade 5:

$$\dot{Y} = \begin{cases} 1 & \text{Hauptschule} \\ 2 & \text{Realschule} \\ 3 & \text{Gymnasium} \end{cases} \tag{7.2}$$

Table 7.1.1 Artificial data for illustration.

		number of cases with		
x	z	$\dot{Y}=1$	$\dot{Y}=2$	$\dot{Y}=3$
0	0	2400	1200	400
0	1	1500	900	600
1	0	800	400	800
1	1	200	100	700

The model assumes that the probabilities $\Pr(\dot{Y}=j)$ depend on values of one or more explanatory variables. In (7.1), there are two variables, $\ddot{X}$ and $\ddot{Z}$, but, of course, more can be added.

For an example, I assume that $\ddot{X}$ is an exogenous explanatory variable that represents some aspect of the student's family background, and $\ddot{Z}$ is a possibly endogenous explanatory variable that represents the student's 'level of academic performance' in the situation where the selection takes place (end of grade 4).

2. Illustration with artificial data

For a simple illustration, I assume that both explanatory variables are binary and that data are given as shown in Table 7.1.1. One can immediately calculate values of the functional relationship

$$(x,z) \longrightarrow \Pr(\dot{Y}=j|\ddot{X}=x,\ddot{Z}=z) \tag{7.3}$$

as shown in the following table:

x	z	$\Pr(\dot{Y}=1\|x,z)$	$\Pr(\dot{Y}=2\|x,z)$	$\Pr(\dot{Y}=3\|x,z)$
0	0	0.6	0.3	0.1
0	1	0.5	0.3	0.2
1	0	0.4	0.2	0.4
1	1	0.2	0.1	0.7

(7.4)

If the number of possible values of the explanatory variables is large, it is often preferable to use a parametric version of the functional relationship (7.3). Often used is a multinomial logit model.

3. The multinomial logit model

This is an extension of the binary logit model that was discussed in Chapter 3. I begin with a version that starts from assuming

$$\Pr(\dot{Y}=j|\ddot{X}=x,\ddot{Z}=z) \approx \frac{1}{A}\exp(\alpha_j + x\beta_{xj} + z\beta_{zj} + xz\beta_{xzj})$$

Since $\sum_j \Pr(\dot{Y}=j|x,z)=1$, it follows that A should be taken as

$$A = \sum\nolimits_j \exp(\alpha_j + x\beta_{xj} + z\beta_{zj} + xz\beta_{xzj})$$

and the basic form of the model becomes

$$\Pr(\dot{Y}=j|\ddot{X}=x,\ddot{Z}=z) \approx \frac{\exp(\alpha_j + x\beta_{xj} + z\beta_{zj} + xz\beta_{xzj})}{\sum_k \exp(\alpha_k + x\beta_{xk} + z\beta_{zk} + xz\beta_{xzk})}$$

However, since probabilities add to unity, it suffices to estimate parameters for $m-1$ alternatives. One therefore adds the constraint that all parameters which relate to the first alternative ($j=1$) are zero: $\alpha_1 = \beta_{x1} = \beta_{z1} = \beta_{xz1} = 0$. The model's basic form then finally is

$$\Pr(\dot{Y}=j|\ddot{X}=x,\ddot{Z}=z) \approx \frac{\exp(\alpha_j + x\beta_{xj} + z\beta_{zj} + xz\beta_{xzj})}{1+\sum_{k=2,m}\exp(\alpha_k + x\beta_{xk} + z\beta_{zk} + xz\beta_{xzk})} \tag{7.5}$$

for $j=2,\ldots,m$. Parameters of the model can be estimated with the maximum likelihood method. To illustrate, I use the data in Table 7.1.1. One gets the following parameter values:

	$j=2$	$j=3$
$\hat{\alpha}_j$	−0.6931	−1.7918
$\hat{\beta}_{xj}$	−0.0000	1.7918
$\hat{\beta}_{zj}$	0.1823	0.8755
$\hat{\beta}_{xzj}$	−0.1823	0.3773

Obviously, one cannot immediately interpret these parameters. They should be used to calculate estimates of the conditional probabilities. This can easily be done by inserting the parameter estimates into (7.5). In our example, since we used a saturated model, one gets the probabilities shown in (7.4). If we had used a model without an interaction term, results would be slightly different.

4. Constraints on available alternatives

An important consideration concerns the available alternatives. While the model presupposes a set of alternatives, $\mathcal{Y}=\{1,\ldots,m\}$, it might well be that the actually available alternatives depend on the circumstances, that is, on values of explanatory variables.

To illustrate, I continue with the example introduced in **7.1.1**. As defined in (7.2), there are three alternatives (school types). I now assume that the actually available alternatives depend on the student's value of $\dot{Z}$,

Box 7.1.1 Data generation for the example.

for $(x = 0, \ldots, 4)$
  for $(i = 1, \ldots, 10000)$
    set $x_i = x$, z_i = random draw from a normal distribution with mean $\mu_{x_i} = 0.05\, x_i$ and variance 1.
    r_i = random number equally distributed in $[0, 1]$.
    (a) if $z_i < -1$, then $y_i = 1$.
    (b) if $-1 \le z_i < 0$, then:
        if $r_i \le \mu_{x_i} + 0.5(z_i + 1)$ then $y_i = 2$, else $y_i = 1$.
    (c) if $z_i \ge 0$, then:
        if $r_i \le \mu_{x_i} + 0.5\,\frac{\exp(z_i)}{1+\exp(z_i)}$ then $y_i = 3$, else $y_i = 2$.

the level of academic performance at the end of grade 4.[1] More specifically, I assume that $\dot{Y} = 1$ is always available, but $\dot{Y} = 2$ requires that $\dot{Z} \ge z_2$, and $\dot{Y} = 3$ requires that $\dot{Z} \ge z_3$ $(z_3 > z_2)$. These constraints must be taken into account in the set-up of the model. This can be done with the help of indicator variables

$$V_2 := I[\dot{Z} < z_2] \quad \text{and} \quad V_3 := I[\dot{Z} < z_3]$$

These variables, together with suitably constrained parameters, can be included into the specification of the multinomial model. For example, omitting the interaction term, the numerator of (7.5) can be specified by

$$\exp(\alpha_j + x\beta_{xj} + v_2\delta_{2j} + v_3\delta_{3j}) \tag{7.6}$$

and the model is estimated with the following constraints:

$$\delta_{23} = \delta_{32} = 0 \quad \text{and} \quad \delta_{22} = \delta_{33} = -\text{LN} \tag{7.7}$$

where LN is a large number (so that the resulting probability is essentially zero).

To illustrate the procedure, I use an extension of the example discussed in Section 6.1. As in that example, there are five levels of parents' SES: $\ddot{X} \in \{0, 1, 2, 3, 4\}$, and $\dot{Z}$ is a normally distributed variable with mean μ_x and variance 1. The mean values depend on x as shown in Table 6.1.1. As threshold values I assume $z_2 = -1$ and $z_3 = 0$.

[1] As indicated by the notation $\dot{Z}$, this is now taken as an endogenous variable having a distribution depending on $\ddot{X}$.

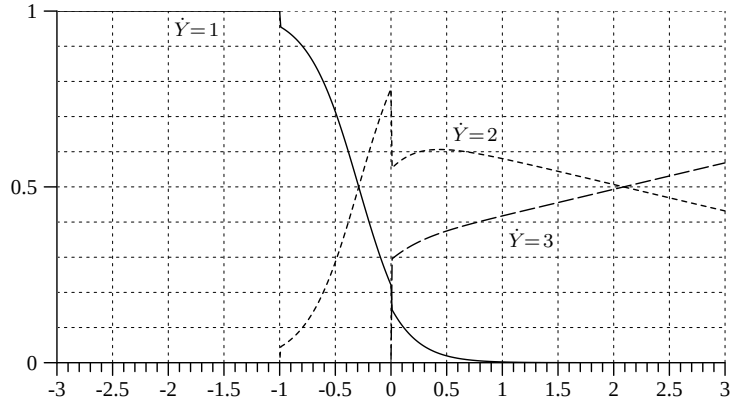


Figure 7.1.1 Dependence of $\Pr(\dot{Y} = j | \ddot{X} = 1, \dot{Z} = z)$ on z (shown on the abscissa). Calculated with (7.9).

For the example, the data are generated as shown in Box 7.1.1. The model is specified as

$$\Pr(\dot{Y} = j | \ddot{X}_0 = x_0, \ldots, \ddot{X}_4 = x_4, \dot{Z} = z) \approx \frac{\exp(x_0\beta_{0j} + \cdots + x_4\beta_{4j} + z\beta_{zj} + v_2\delta_{2j} + v_3\delta_{3j})}{1 + \sum_{k=2,m} \exp(x_0\beta_{0k} + \cdots + x_4\beta_{4k} + z\beta_{zk} + v_2\delta_{2k} + v_3\delta_{3k})} \tag{7.8}$$

with five dummy variables $\ddot{X}_j := I[\ddot{X} = j]$ $(j = 0, \ldots, 4)$. One finds the following parameter estimates:[2]

	$j = 2$	$j = 3$
$\hat{\beta}_{0j}$	1.0041	0.1966
$\hat{\beta}_{1j}$	1.2606	0.6287
$\hat{\beta}_{2j}$	1.3743	0.9182
$\hat{\beta}_{3j}$	1.6734	1.4128
$\hat{\beta}_{4j}$	1.8079	1.7520
$\hat{\beta}_{zj}$	4.3503	4.6527
$\hat{\delta}_{2j}$	−20.0000	0.0000
$\hat{\delta}_{3j}$	0.0000	−20.0000

(7.9)

These parameters can be used to calculate estimates of the conditional probabilities defined by the model. How they depend on values of $\dot{Z}$ is shown in Figures 7.1.1 and 7.1.2 for $\ddot{X} = 1$ and $\ddot{X} = 4$, respectively.

[2] Estimation was done with the `qreg` procedure of the program `TDA` that allows one to specify parameter constraints. LN was set to 20.

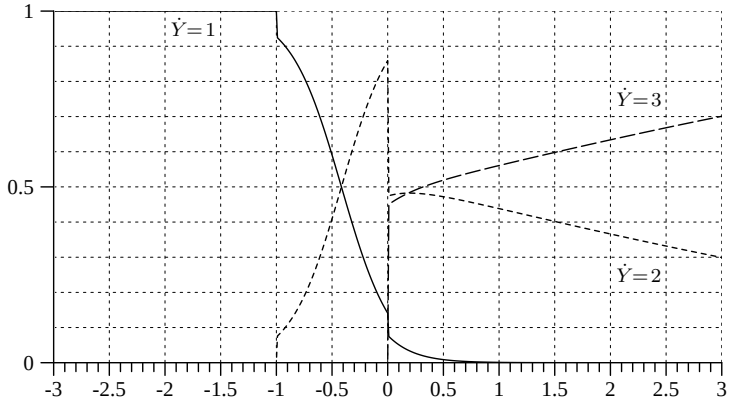


Figure 7.1.2 Dependence of $\Pr(\dot{Y}=j|\ddot{X}=4,\dot{Z}=z)$ on z (shown on the abscissa). Calculated with (7.9).

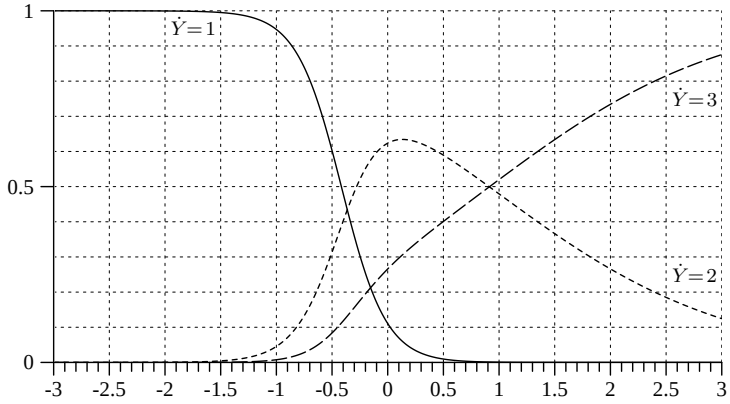


Figure 7.1.3 Dependence of $\Pr(\dot{Y}=j|\ddot{X}=4,\dot{Z}=z)$ on z (shown on the abscissa). Calculated with (7.10).

5. Relevance of observing constraints

Constraints on available alternatives must explicitly be taken into account when specifying a model. In order to illustrate the relevance, I reestimate the model without constraints by simply omitting the variables v_2 and v_3 from the specification (7.8). One then gets the following parameter estimates:

	$j=2$	$j=3$
$\hat{\beta}_{0j}$	1.0004	−0.5081
$\hat{\beta}_{1j}$	1.2467	−0.1169
$\hat{\beta}_{2j}$	1.3442	0.1411
$\hat{\beta}_{3j}$	1.6238	0.5901
$\hat{\beta}_{4j}$	1.7296	0.8770
$\hat{\beta}_{zj}$	4.7664	5.7005

(7.10)

They are obviously quite different from the values in (7.9). In particular, they suggest a misleading impression of the dependence on $\dot{Z}$, the level of academic performance. This is also shown in Figure 7.1.3 which is based on the parameter estimates (7.10).

7.2 School type selection in grade 5

In this section, I use data from the the AIDA survey (*Aufwachsen in Deutschland: Alltagswelten*), conducted by the *Deutsches Jugendinstitut* (Munich), to illustrate a multinomial logit model of the selection of school types in grade 5. I thank Ulrich Pötter who kindly provided the data.

1. Data from the AIDA survey

I consider children being in grade 5 at the interview date (second half of 2009), altogether 592 children (AIDA variable: `V1_B24`). 46 of these children are still in a primary school (Grundschule). The distribution on school types of the 546 children who already changed to a secondary level (AIDA variable `V1_B25A`) is as follows:

cases	school type
47	Hauptschule
102	Realschule
288	Gymnasium
69	Gesamtschule
40	Sonstige

I consider the 506 children attending one of the first four school types.

Table 7.2.1 Distribution of marks attained in grade 4 (-1 means 'not available').

	mark in *Mathematik*						mark in *Deutsch*					
$\dot{Y}$	1	2	3	4	5	-1	1	2	3	4	5	-1
1	0	7	22	13	0	5	0	9	21	11	1	5
2	2	39	48	8	2	3	1	33	57	8	1	2
3	63	143	68	11	0	3	57	160	59	10	0	2
4	6	19	26	5	0	13	4	26	22	9	0	8
	71	208	164	37	2	24	62	228	159	38	2	17

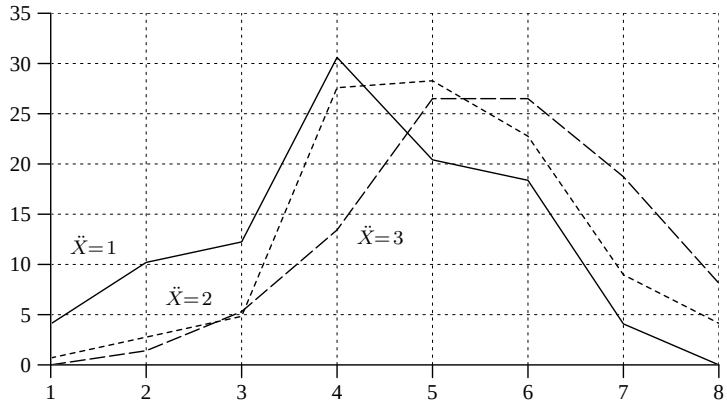


Figure 7.2.1 Distribution of the index defined in (7.13) conditional on parents' educational level. The mean values are 4.24, 5.05, and 5.60, respectively.

Correspondingly, I use the following dependent variable:

$$\dot{Y} = \begin{cases} 1 & \text{Hauptschule} \\ 2 & \text{Realschule} \\ 3 & \text{Gymnasium} \\ 4 & \text{Gesamtschule} \end{cases} \tag{7.11}$$

As explanatory variables, I use the parents' educational level, and an index of the marks attained in grade 4. Table 7.2.1 shows the distribution of these marks in the subjects *Mathematik* and *Deutsch* (AIDA variables `V1_B21A` and `V1_B21B`).

To record parents' educational level (based on the AIDA variable `HBIEL`) I distinguish three levels:

$$\ddot{X} = \begin{cases} 1 & \text{Hauptschule or without graduation} \\ 2 & \text{Realschule} \\ 3 & \text{Abitur, Fachabitur} \end{cases} \tag{7.12}$$

In the following, I use 477 cases with known values of $\ddot{X}$ and the marks shown in Table 7.2.1. In order to include these marks into the model, I use an index

$$\dot{Z} = 10 - \text{mark\_in\_Mathematik} - \text{mark\_in\_Deutsch} \tag{7.13}$$

Figure 7.2.1 shows distributions of this index conditional on parents' educational level.

2. A model of school type selection

I now consider a functional model with $\dot{Y}$ as an endogenous variable, and $\ddot{X}$ and $\dot{Z}$ as explanatory variables. I assume that childrens' marks depend (to some degree) on their parents' educational level and thus consider $\dot{Z}$ as an endogenous explanatory variable. This allows one to think of total effects of the variable $\ddot{X}$ (see below).

As a parametric form of the model I use a multinomial logit model and represent $\ddot{X}$ by three dummy variables, $\ddot{X}_l := I[\ddot{X} = l]$:

$$\Pr(\dot{Y}=j|\ddot{X}_1=x_1, \ddot{X}_2=x_2, \ddot{X}_3=x_3, \dot{Z}=z) \approx \frac{\exp(x_1\beta_{1j} + x_2\beta_{2j} + x_3\beta_{3j} + z\beta_{zj})}{1 + \sum_{k=2,m} \exp(x_1\beta_{1k} + x_2\beta_{2k} + x_3\beta_{3k} + z\beta_{zk})} \tag{7.14}$$

Estimating the model with the maximum likelihood method, one finds the following parameter values (standard errors in brackets):

	$j=2$	$j=3$	$j=4$
$\hat{\beta}_{1j}$	−1.8148 (0.71)	−5.8740 (0.86)	−3.4528 (0.87)
$\hat{\beta}_{2j}$	−0.8952 (0.73)	−4.2216 (0.80)	−2.5260 (0.86)
$\hat{\beta}_{3j}$	−0.6554 (0.75)	−3.1038 (0.80)	−2.3337 (0.88)
$\hat{\beta}_{zj}$	0.4585 (0.16)	1.2167 (0.17)	0.6868 (0.18)

(7.15)

These parameter values can be used to calculate estimates of the probabilities of changing into one of the four possible school types, and one can investigate how these probabilities depend on values of the explanatory variables. This will be done below.

A further question concerns how well the model fits the data. One possible approach is based on using the model for individual predictions. For each individual case, i, one can calculate values $\hat{p}_{ij}$, that is, the estimated probability that i was selected into the school type j. Then one

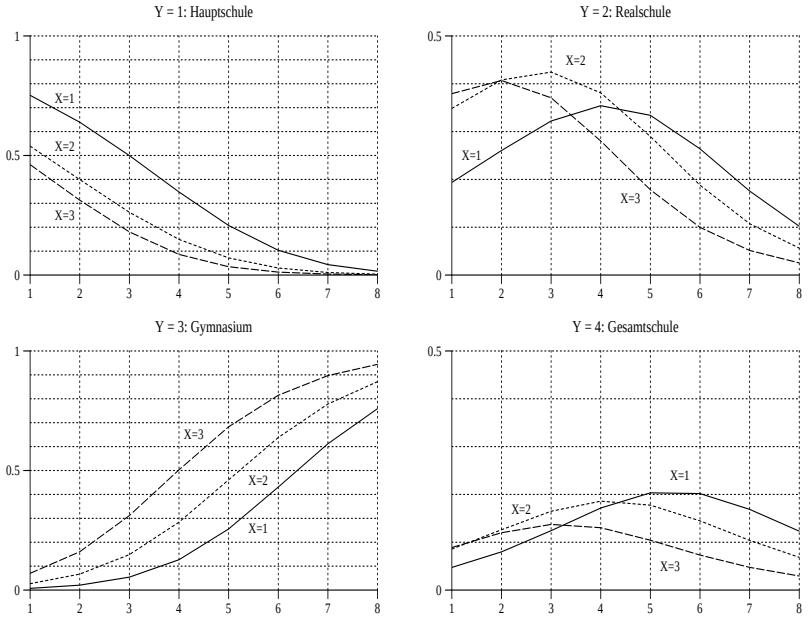


Figure 7.2.2 Dependence of $\Pr(\dot{Y}=j|\ddot{X}=x,\dot{Z}=z)$ on values of $\dot{Z}$ (index of marks), for different values of $\ddot{X}$ (parents' educational level).

can predict that i will be selected into the school type having the highest probability. Finally, one can compare these predicted with the actually observed school types. Based on the parameter estimates in (7.15), one finds that $302/477 = 63\,\%$ can be correctly predicted.

3. Effects of explanatory variables

I now consider how the estimated probabilities depend on values of the explanatory variables. A first answer is given by the plots in Figure 7.2.2. It is seen that both variables, parents' educational level and childrens' marks, substantially influence the selection of school types.

Assuming that childrens' marks depend (to some degree) on the parents' educational level, one can distinguish between primary and secondary effects. As was discussed in **6.1.4**, one can interpret

$$H_j(x,x') := \sum\nolimits_z \Pr(\dot{Y}=j|\ddot{X}=x',\dot{Z}=z)\Pr(\dot{Z}=z|\ddot{X}=x) \qquad (7.16)$$

as the probability of $\dot{Y}=j$ under the assumption that the primary effect is given by x and the secondary effect is given by x'. Analogous to (6.1), one

Table 7.2.2 Decomposition into primary and secondary effects according to (7.16) and (7.17).

			Total effect	Primary effect	Secondary effect
j	x'	x''	$H_j(x'',x'')-H_j(x',x')$	$H_j(x'',x'')-H_j(x',x'')$	$H_j(x',x'')-H_j(x',x')$
1	1	2	-0.030	-0.010	-0.020
1	1	3	0.052	0.174	-0.122
1	2	3	0.082	0.084	-0.002
2	1	2	0.115	-0.047	0.162
2	1	3	0.394	0.199	0.195
2	2	3	0.279	0.086	0.193
3	1	2	0.079	-0.057	0.136
3	1	3	0.619	0.191	0.428
3	2	3	0.540	0.071	0.469

can then use the decomposition

$$H_j(x'',x'') - H_j(x',x') = \qquad (7.17)$$
$$[H_j(x'',x'') - H_j(x',x'')] + [H_j(x',x'') - H_j(x',x')]$$

where the first term on the right-hand side represents the primary effect, and the second term represents the secondary effect.

Calculation of the quantities $H_j(x,x')$ uses the estimated model parameters (for the first term on the right-hand side), and the observed distributions of the index of marks as shown in Figure 7.2.1 and, numerically, in the following table:

z	$\Pr(\dot{Z}=z\|\ddot{X}=1)$	$\Pr(\dot{Z}=z\|\ddot{X}=2)$	$\Pr(\dot{Z}=z\|\ddot{X}=3)$
1	0.0408	0.0069	0.0000
2	0.1020	0.0276	0.0141
3	0.1224	0.0483	0.0530
4	0.3061	0.2759	0.1343
5	0.2041	0.2828	0.2650
6	0.1837	0.2276	0.2650
7	0.0408	0.0897	0.1873
8	0.0000	0.0414	0.0813

Results are shown in Table 7.2.2. It is seen that the secondary effects are quite large and often much more important than the primary effects.

4. Differences between boys and girls?

In order to investigate differences between boys and girls, I add a binary variable, $\ddot{S}$, to the model which takes the value 1 for girls and 0 for boys.

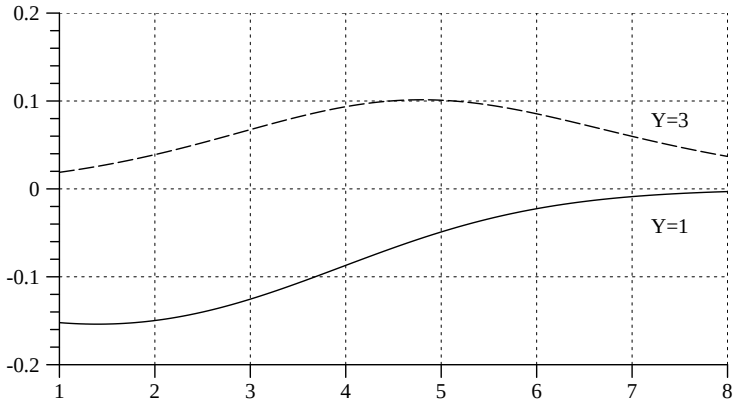


Figure 7.2.3 Plot of the functions (7.19), calculated for $\ddot{X}=2$.

The corresponding model parameters are β_{sj}. Parameter estimates and standard errors for this enlarged model are shown in the following table:

	$j=2$	$j=3$	$j=4$
$\hat{\beta}_{1j}$	−2.1904 (0.77)	−6.4268 (0.91)	−3.8978 (0.92)
$\hat{\beta}_{2j}$	−1.1312 (0.75)	−4.5962 (0.82)	−2.8049 (0.88)
$\hat{\beta}_{3j}$	−0.8595 (0.76)	−3.4341 (0.81)	−2.5791 (0.90)
$\hat{\beta}_{zj}$	0.4660 (0.16)	1.2134 (0.17)	0.6900 (0.18)
$\hat{\beta}_{sj}$	0.5648 (0.42)	0.9489 (0.42)	0.7060 (0.46)

(7.18)

One finds remarkable effects of $\ddot{S}$ for school type 3 (Gymnasium) and, correspondingly, for type 1 (Hauptschule). In order to visualize these effects, I consider the functions

$$z \longrightarrow \Pr(\dot{Y}=j|\ddot{X}=x,\dot{Z}=z,\ddot{S}=1)-\Pr(\dot{Y}=j|\ddot{X}=x,\dot{Z}=z,\ddot{S}=0) \quad (7.19)$$

Figure 7.2.3 shows these functions for $j=1$ and $j=3$, with $\ddot{X}=2$.

Chapter 8

Causal interpretations

The functional relationships which are posited when specifying a functional model do not automatically have a causal meaning. In general, their presupposition only entails the claim that variables used as arguments in a function can contribute to predicting conditional distributions of the dependent variable. For example, one can use a person's educational level to predict the educational level of her parents; but there obviously is no corresponding causal relationship. So the question arises in which sense one sometimes can claim that relationships in a functional model also have a causal meaning.

This question concerns the understanding of causal relationships and must be distinguished from the further question of how one can use statistical data for estimating quantitative (numerically specified) forms of such relationships. In the present chapter, I am mainly concerned with the conceptual question. The first section proposes a rudimentary understanding of the idea that variables can be causally related. In the second section, I introduce a distinction between explanatory and treatment models, and

suggest that there correspondingly are some differences in the understanding of causal effects. The third section takes up the often used notion of 'potential outcomes' and shows that there are two quite different understandings.

8.1 Causal relations between variables

1. Causally relevant variables

Consider a functional relationship

$$x \longrightarrow \Pr[\dot{Y}|\ddot{X}=x]$$

What is entailed by the idea that this can be viewed as a causal relationship? I propose that there are basically two claims:

(1) that one can refer to a fact-generating process generating values of $\dot{Y}$,[1] and

(2) that this process depends on values of $\ddot{X}$.

I then call $\ddot{X}$ a variable which is *causally relevant* for $\dot{Y}$.

These two claims are at the core of viewing causation as a generative process (Goldthorpe 2001, Blossfeld 2009). Of course, they also provide a starting point for further questions: How do processes generating values of $\dot{Y}$ depend on $\ddot{X}$, and which role is played by further variables on which $\dot{Y}$ possibly depends?

2. Thinking of causes as events

In order to think about how processes generating values of $\dot{Y}$ depend on $\ddot{X}$ one must be more specific about the meaning of the values of $\ddot{X}$. I begin with assuming that values of $\ddot{X}$ represent possible events. In the most simple case, $\ddot{X}$ is a binary variable, and $\ddot{X}=1$ means that an event of a specified kind has occurred, and $\ddot{X}=0$ means that such an event has not (yet) occurred.

One can then distinguish two kinds of relationships between $\ddot{X}$ and a process generating values of $\dot{Y}$:

a) The event initiates a process that generates a value of $\dot{Y}$. This presupposes that the occurrence of the event is a necessary condition for the generation of a value of $\dot{Y}$. As an example, think of $\dot{Y}$ as the outcome of a student's participating in a learning frame. The student's *beginning* to participate in the learning frame can be considered as an event that initiates a process that eventually generates a value of $\dot{Y}$.

[1] See the distinction between data-generating and fact-generating processes introduced in § 2.1.2.

b) The event occurs while a process that eventually generates a value of $\dot{Y}$ already takes place. So one can think of two such processes: one during which the event did occur, and another one in which the event did not occur. The impact of the event, if it occurs, must then be understood as modifying an ongoing process. As an example, one can think that the student becomes severely ill while participating in the learning frame.

Most often, already the definition of a process requires to refer to an event that initiates the process. The causal relevance of that event is then easily stated: Its occurrence is a necessary condition for the process to take place.

Variables representing the occurrence of events will be called *event variables* (Rohwer 2010, ch. 7). These need not be binary variables which refer to just one event type. In general, if $\dot{X}$ is an event variable, its domain will be denoted by

$$\mathcal{X} = \{0, 1, \ldots, m\}$$

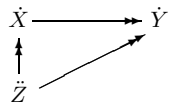
with values having the following meaning: If $x > 0$, $\dot{X}=x$ means that an event of the type x has occurred; $\dot{X}=0$ means that no event of the specified kinds has yet occurred. The definition shows that using event variables at least implicitly requires a temporal view.

3. Causally relevant conditions

Causally relevant variables need not be event variables. As another kind one can think of variables representing conditions on which a process depends. I then speak of *context variables*.

To illustrate, I use our standard example in which a student's educational success, $\dot{Y}$, depends on the school type, $\dot{X}$, and on the parents' educational level, $\ddot{Z}$. The functional model looks as follows:

Model 8.1.1



In this model, one can think of $\dot{X}$ as a variable representing the occurrence of an event:

$$\dot{X} = \begin{cases} 1 & \text{if the student starts participating in learning frame } \sigma_1 \\ 2 & \text{if the student starts participating in learning frame } \sigma_2 \end{cases}$$

In contrast, $\ddot{Z}$, recording the parents' educational level, is a context variable. Its values can sensibly be understood as characterizing the context in which the process that generates a value of $\dot{X}$ takes place. Similarly, one

can understand the causal relevance of $\ddot{Z}$ for processes generating values of $\dot{Y}$. The leading idea is that parents' activities through which they support a child's education depend on their own educational level.

How to think of the causal relevance of $\dot{X}$? There are two considerations. First, thinking of the event $\dot{X} = j$, it can be understood as initiating a process that eventually generates a value of $\dot{Y}$. In this view, $\dot{X}$ is causally relevant because without $\dot{X}$'s taking a positive value a process generating a value of $\dot{Y}$ cannot take place. Second, as soon as one of the possible events did occur, it can be viewed as having generated a specific context. If $\dot{X} = j$, it is the context σ_j, a particular learning frame, in which the process generating a value of $\dot{Y}$ takes place.

This can be generally stated: As soon as an event variable has a positive value, the variable can be considered as a context variable for a process that begins at the point in time when the event occurred.

4. Comparative and dynamic effects

To speak of a causally relevant variable, say $\dot{X}$, presupposes a functional model in which the variable has a particular place and can be functionally related to other variables representing possible effects. However, it is not the model, understood as a system of mathematical functions, that provides the causal meaning. To give a variable a causal meaning requires considerations which cannot be expressed in terms of mathematical functions.

The functional model is silent about the meaning of its functional relationships. But given a causal interpretation, it can be used to formally define causal effects. One specifies a variable, say $\dot{Y}$, representing the outcomes of interest, and considers all variables on which $\dot{Y}$ functionally depends. Assume that these are the variables $\dot{X}$ and $\dot{Z}$. Both can then be used to define effects. For example, an effect of $\dot{X}$ can be defined by

$$\Delta^s(\dot{Y}; \dot{X}[x', x''], \dot{Z}=z) = \\ \mathrm{E}(\dot{Y}|\dot{X}=x'', \dot{Z}=z) - \mathrm{E}(\dot{Y}|\dot{X}=x', \dot{Z}=z) \tag{8.1}$$

This definition compares the expectation of $\dot{Y}$ in two situations: one in which $\dot{X}=x'$ and another one in which $\dot{X}=x''$, and further presupposes that $\dot{Z} = z$ in both situations. In this sense, the definition formulates a *comparative effect*, and can be used for all kinds of causally relevant variables. If $\dot{X}$ is an event variable, also another effect definition becomes possible:

$$\Delta^d(\dot{Y}; \dot{X}[j], \dot{Z}=z) = \mathrm{E}(\dot{Y}|\dot{X}=j, \dot{Z}=z) - \mathrm{E}(\dot{Y}|\dot{X}=0, \dot{Z}=z) \tag{8.2}$$

This definition formulates a *dynamic effect*; it compares a situation where the event $\dot{X} = j$ occurred with a situation where no event (of the kinds specified by $\dot{X}$) occurred.[2]

To illustrate, consider the example introduced in § 8.1.3. A comparative effect compares the educational outcomes in the two learning frames, σ_1 and σ_2. In contrast, a dynamic effect compares the outcome of $\dot{X}=j$ with $\dot{X}=0$, and since a positive value of $\dot{X}$ is a necessary condition for a value of $\dot{Y}$, the dynamic effect is given by $\mathrm{E}(\dot{Y}|\dot{X}=j, \dot{Z}=z)$.

5. Functional and causal mechanisms

Without presupposing possible effects one cannot think of 'causes'. In a statistical approach to causality, possible effects are conceptualized by an outcome variable, $\dot{Y}$. In social research, being interested in processes generating values of $\dot{Y}$, it is seldom reasonable to consider only a single causal condition, say $\ddot{X}$. In most applications one has to take into account further causally relevant conditions. The question 'What is the causal effect of $\ddot{X}$ on $\dot{Y}$?', without further qualification, is then not appropriate.

An alternative is to think of 'mechanisms'.[3] Here I use the term in this sense: A *mechanism* is an explicitly defined framework for thinking of processes generating values of an outcome variable. More specifically, I use the term *functional mechanism* to denote a functional model that shows how an outcome variable depends on other variables; and it will be called a *causal mechanism* if at least some of the functional relationships can be given a causal interpretation.

Given these definitions, a mechanism is a (formal) framework and must be distinguished from the processes which, possibly, take place according to the rules of the mechanism. This entails that a mechanism is not by itself a dynamic entity. While one can sensibly think that a process can generate an outcome, this cannot be said of a mechanism. But note that only the mechanism has an explicit representation (in terms of variables). To think of processes that actually generate outcomes requires a causal interpretation of the mechanism.

8.2 References to human actors

1. Primary and secondary actors

Models in social research most often concern processes which depend on the behavior of human actors. In the following, I call these the *primary actors*. In contrast, I speak of *secondary actors* when referring to those

[2] If the occurrence of an event specified by $\dot{X}$ is a necessary condition for $\dot{Y}$'s getting a value, I use the convention that $\mathrm{E}(\dot{Y}|\dot{X}=0, \dot{Z}=z) = 0$; see Rohwer (2010, ch. 7).

[3] For a discussion of many of the meanings which are given to this term in the sociological literature, see Mahoney (2001).

who construct and use models.

For example, think of the models dealing with students' educational outcomes in different learning frames. Primary actors are the students, their parents, the teachers; in general, all actors to which one refers when interpreting the models and reflecting about causal relationships. In contrast, the secondary actors are those who construct, discuss, and possibly use these models for one reason or another.

Note that the distinction presupposes the reference to a model. Only w.r.t. a model can one distinguish primary and secondary actors in the proposed sense.

2. Explanatory and treatment models

There are many reasons why one could be interested in models. Here I want to mention just one distinction that also suggests to distinguish two kinds of models.

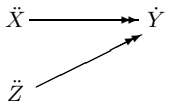
One interest concerns the primary actors, the conditions and outcomes of their behavior. Models are then constructed as tools for understanding and explaining conditions and outcomes of the behavior of the primary actors. I then speak of *explanatory models*.

In contrast, secondary actors could be interested in the possibility of interventions supporting their economic and/or political goals. Models are then constructed as tools for assessing the possible effects of treatments. I then speak of *treatment models*.

3. Functional models of experiments

Discussions of causally interpretable models often presuppose an interest in effects of treatments. Grounded in a long tradition, treatment models are preferably related to an experimental context. I briefly mention some ways in which functional models can be used as a formal framework. A first possibility is illustrated by the following model.

Model 8.2.1

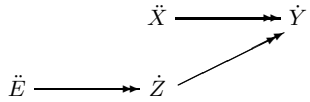


Values of $\ddot{X}$ represent the treatments whose causal effects are of primary interest. $\ddot{Z}$ records further conditions which, presumably, are causally relevant for $\dot{Y}$. Both are exogenous variables because their values are deliberately fixed by the experimenter.

In another kind of experiment, the experimenter randomly selects some of the conditions for the experiment but still deliberately generates values

of the treatment variable. This can be represented by the following model.

Model 8.2.2



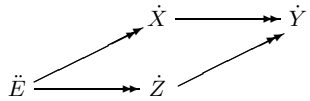
$\ddot{X}$ is still an exogenous variable, but $\dot{Z}$ is now an endogenous variable which depends in a specified way on $\ddot{E}$, an event variable initiating the experiment.

In both situations, assuming that the experiment concerns processes depending on the behavior of primary actors, the experimenter is a secondary actor. There also is then a potential conflict. Being interested in predictable effects of treatments, the experimenter has reason to control and regulate the behavior of the primary actors as far as possible.

4. Randomly assigned causal conditions

There is a further kind of experiment in which also values of the treatment variable, $\ddot{X}$, are randomly generated. The functional model then looks as follows:

Model 8.2.3



The set-up entails that, conditional on $\ddot{E}=1$ (initiation of the experiment), $\dot{X}$ and $\dot{Z}$ are independent. In a standard notation, this can be written as

$$\dot{X} \perp\!\!\!\perp \dot{Z} \,|\, \ddot{E} = 1$$

This allows one to express effects of $\dot{X}$ as follows:

$$\Delta^s(\dot{Y}; \dot{X}[x', x''], \ddot{E}=1) = \sum\nolimits_z \Delta^s(\dot{Y}; \dot{X}[x', x''], \dot{Z}=z, \ddot{E}=1) \Pr(\dot{Z}=z|\ddot{E}=1) \tag{8.3}$$

This shows that a randomized experiment allows one to compare effects of different treatments, values of $\dot{X}$, in a balanced way: the distribution of further possibly relevant variables, $\dot{Z}$, does not depend on the values of $\dot{X}$, the treatments, that one intends to compare. But note that this does not entail that effects of $\dot{X}$ are independent of $\dot{Z}$. If $\dot{X}$ and $\dot{Z}$ interact in the generation of values of $\dot{Y}$, the effect defined in (8.3) still depends on the distribution of $\dot{Z}$.

Notice that a reference to the variable $\ddot{E}$ is required in order to think of an experimental context which includes an experimenter who at least initiates the experiment. Without the variable $\ddot{E}$, Model 8.2.3 only contains random variables getting their values from processes not represented in the model. Also note that this is an intervention model in a very specific sense: randomization is used as a fact-generating process, that is, a process generating events (values of $\dot{X}$).

5. Two contexts for randomization

Is randomization useful in social research? In thinking about this question one should distinguish between two different contexts for randomization.

First, one can think of data-generating processes. The most relevant application concerns the selection of units (for further observation). There are good arguments that samples should be generated randomly, that is, according to in some way fixed and known selection probabilities.

An essentially different context is experimentation. Performing an experiment requires, first of all, a fact-generating process. In this context, randomization not only, if at all, concerns the selection of a sample of units for the experiment; but also is a method of creating facts ('treatments') whose possible effects one intends to study.

While randomization in data generation is certainly useful in social research, the value of randomized experiments seems questionable. The argument is not that such experiments are seldom possible. The relevant point is that randomization would change, in an essential way, the processes to be studied.

The example depicted in Model 8.1.1 can show this. In this example, one can be interested in effects of the learning frames ($\dot{X}$). The model realistically assumes that the selection of learning frames depends on the student's family background (represented by $\ddot{Z}$). This is a fact-generating process. Randomization would substitute this by another fact-generating process that randomly assigns students to learning frames.

8.3 Rule-based and descriptive approaches

In the statistical literature, discussions of causal effects often use notions of 'potential outcomes' (e.g. Rubin 2005; Morgan and Winship 2007; Angrist and Pischke 2009; Gangl 2010). Different understandings are possible. In this section I distinguish, and contrast, rule-based and descriptive understandings of potential outcomes.

1. Rule-based understanding of potential outcomes

The approach to understanding causal relationships that was sketched in Section 8.1 uses functional models as a formal framework. This is appropriate when one is interested in causally interpretable rules. Such rules – I briefly speak of *causal rules* – concern potential outcomes which can be linked to different values of causally relevant variables. Continuing with the notation introduced in Section 8.1, such a rule has the form

$$(x, z) \longrightarrow \mathrm{E}(\dot{Y}|\dot{X}=x, \dot{Z}=z) \tag{8.4}$$

It is a causal rule if $\dot{X}$ and $\dot{Z}$ can be interpreted as variables which are causally relevant for the generation of values of $\dot{Y}$. The dependent variable, $\dot{Y}$, represents *potential outcomes* which can be expected in a generic situation where $\dot{X}$ and $\dot{Z}$ have specified values. Given the rule, one can compare potential outcomes for different values of $\dot{X}$ and $\dot{Z}$, and, as defined in (8.1) and (8.2), interpret their differences as causal effects.

2. Descriptive notion of potential outcomes

In the statistical literature, one also finds a descriptive notion of potential outcomes which is not based on rules. This notion relates to a specified set of particular units, say Ω, and presupposes three (or more) statistical variables.[4] A variable

$$X : \Omega \longrightarrow \mathcal{X}$$

represents causally relevant factors (conditions or events). As is often done in the literature, in order to simplify notations, I assume that X is a binary variable:

$$X = \begin{cases} 1 & \text{if a specified causal factor is present} \\ 0 & \text{otherwise} \end{cases}$$

The interest concerns outcomes in situations where $X = 1$ or $X = 0$. It is assumed that these outcomes can be represented by statistical variables

$$Y_j : \Omega \longrightarrow \mathcal{Y} \tag{8.5}$$

($j = 0$ or 1) having the following meaning: If $X(\omega) = j$, the outcome of interest has the value $Y_j(\omega)$.

One can then formally define, for each unit $\omega \in \Omega$, a causal effect $Y_1(\omega) - Y_0(\omega)$. Of course, these individual causal effects cannot be observed. One therefore aims to estimate an *average causal effect* which can be defined for Ω by

$$\mathrm{M}(Y_1) - \mathrm{M}(Y_0) \tag{8.6}$$

[4] See, for example, Holland (1986). Similarly, Rubin (2005: 323) refers to an array of given, partly observed values.

However, observations only allow estimation of conditional mean values, $\mathrm{M}(Y_j|X=j)$. So the question arises under which conditions one can think of these conditional mean values as unbiased estimates of $\mathrm{M}(Y_j)$. A sufficient condition would be that X and Y_j are approximately independent;[5] formally: $Y_j \perp\!\!\!\perp X$ (for $j = 0, 1$).

This independence condition suggests that a critical question concerns the generation of values of X. If possible, such values should be randomly assigned to the members of Ω. This would justify to consider $\Omega_j := \{\omega \in \Omega | X(\omega) = j\}$ as a simple random sample from Ω, and therefore to assume that Y_j has approximately the same distribution in Ω_j and Ω.

3. Including further causally relevant variables

As described in the previous paragraph, the descriptive approach aims to define a causal effect that can be attributed to a single variable, X. A somewhat extended formulation is required if effects also depend on further variables. Assume that outcomes depend not only on X, but also on values of a variable Z (possibly consisting of several components). Instead of (8.6), one has to consider the effect definition

$$\mathrm{M}(Y_1|Z=z) - \mathrm{M}(Y_0|Z=z) \tag{8.7}$$

As before, values of Y_j can only be observed if $X = j$; the observable conditional mean values are $\mathrm{M}(Y_j|X = j, Z = z)$. They provide unbiased estimates of $\mathrm{M}(Y_j|Z=z)$ if

$$Y_j \perp\!\!\!\perp X \,|\, Z = z \quad (\text{for } j = 0, 1) \tag{8.8}$$

This shows that it would suffice to perform the randomization (the random assignment of values of X to the members of Ω) separately for each value of Z.

4. Balanced effects and kinds of models

Neither the rule-based nor the descriptive approach require that the explicitly represented variables, X and Z, are independent. Relationships between these variables become important, however, if one aims to define causal effects of just one variable, say X. Whether this is possible depends first of all on whether X and Z interact in the generation of outcomes.[6] If they interact, effects cannot be attributed solely to X. It is nevertheless possible to define average effects. Following the rule-based approach, one can use the definition

$$\sum\nolimits_z \big(\mathrm{E}(\dot{Y}|\dot{X}=x'', \dot{Z}=z) - \mathrm{E}(\dot{Y}|\dot{X}=x', \dot{Z}=z)\big)\,\mathrm{Pr}(\dot{Z}=z) \tag{8.9}$$

where $\mathrm{Pr}(\dot{Z}=z)$ refers to an (arbitrarily) specified distribution of $\dot{Z}$. This is a *balanced effect*, meaning that the distribution of $\dot{Z}$ is identical for x' and x''. Of course, if $\dot{X}$ and $\dot{Z}$ interact, the effect still depends on the assumed distribution of $\dot{Z}$.

How to proceed when following the descriptive approach depends on the given data to which the causal statements relate. If the data result from a process which entailed a randomization of X w.r.t. Z, the distribution of Z already is approximately independent of X, and one can interpret (8.6) as a balanced average effect of X. Again, if X and Z interact, this effect also depends on the distribution of Z in the reference set of units.

If X and Z are not independent, one can construct a balanced effect. This is analogous to the procedure in the rule-based approach. In the descriptive approach, one starts from (8.7) and (arbitrarily) specifies a distribution of Z. Formally analogous to (8.9), an average effect can then be defined by

$$\sum\nolimits_z \big(\mathrm{M}(Y_1|Z=z) - \mathrm{M}(Y_0|Z=z)\big)\,\mathrm{P}(Z=z) \tag{8.10}$$

One might ask whether balanced effects are particularly useful. This depends on the kind of model. With treatment models, one is normally interested in finding an effect that can be attributed solely to the treatment, given that all other possibly relevant conditions are in some sense fixed. This interest suggests to construct balanced effects.

The situation is different with explanatory models. In social research, explanatory models most often relate to situations where at least some of the causally relevant conditions are generated by actions of primary agents. Effects of single variables are then never balanced w.r.t. all causally relevant conditions. It would be possible, of course, to construct balanced effects w.r.t. observed variables; but I think that a primary interest concerns how the real effects, which are unbalanced, come into being.

5. Contrasting the two approaches

The rule-based and the descriptive approach to the definition of causal effects are in several respects different.

(1) A first difference concerns the notion of potential outcomes. As mentioned in § 8.3.1, the rule-based approach conforms to the understanding that potential outcomes are outcomes which, under specified conditions, possibly will come into existence. Correspondingly, potential outcomes are

[5] Note that X and Y_j are statistical variables, defined for a finite reference set Ω. One can therefore think of statistical independence only in an approximate sense.

[6] As was discussed in Chapter 4, this must be distinguished from dependency relations which concern the joint distribution of X and Z (or $\dot{X}$ and $\dot{Z}$).

defined by a rule (a linguistic if-then construction).

The descriptive approach, in contrast, presupposes that potential outcomes (= values of Y_0 and Y_1) already exist before values of X, and other causally relevant variables, are fixed. To speak of 'potential outcomes' is therefore somewhat misleading. Actually, what is potentially realized is an observation of an already existing fact (value of Y_j).[7] So it would be less confusing to speak of 'potential observations'.

(2) It might be helpful to remember the distinction between fact-generating and data-generating processes. The rule-based approach aims to formulate causal rules for fact-generating processes. The descriptive approach, as it is theoretically formulated, is concerned with data-generating processes which provide partial information about hypothetically presupposed facts (values of Y_0 and Y_1). This approach therefore seems to allow one to think of 'causal inference' in parallel to a missing observation problem.[8]

(3) It is important to understand that the variables Y_0 and Y_1 can only be defined by referring to a set of existing units. For each particular unit, say $\omega \in \Omega$, one can posit values, $Y_0(\omega)$ and $Y_1(\omega)$, representing the outcomes corresponding to $X(\omega) = 0$ and $X(\omega) = 1$, respectively. This is possible because, and insofar, one can assume that all further conditions on which outcomes depend are implicitly fixed by the reference to ω, a particular unit existing in particular circumstances.[9] The descriptive approach is therefore essentially static and not well suited for causal interpretations of temporally extended processes.

The rule-based approach, in contrast, is not based on a reference to a set of already existing units, but relates to generic units which are only defined by values of variables. It is therefore not possible to define variables corresponding to Y_0 and Y_1. Instead, there is a single outcome variable, $\dot{Y}$, having possible values which only become realized when, and after, $\dot{X}$, and any further variables which define the generic unit, have taken specific values. There are no restrictions for thinking of a temporally extended process connecting $\dot{X}$ and the final outcome, $\dot{Y}$.

(4) As a consequence, the independence requirement (8.8) cannot be formulated in the conceptual framework of the rule-based approach. Of course, based on the mentioned understanding of $\dot{Y}$, one can define variables $\dot{Y}_j$ having distributions defined by $\Pr[\dot{Y}_j|\dot{Z} = z] = \Pr[\dot{Y}|\dot{X} = j, \dot{Z} = z]$. An

[7] This is seldom explicitly mentioned; but see Greenland (2004: 4).

[8] See, e.g., Rubin et al. (2004: 105), Winship and Morgan (1999: 664).

[9] Positing values of Y_0 and Y_1 can be done in a deterministic or in a probabilistic way. This entails different understandings of individual causal effects, but the essential features of the descriptive approach are independent of this distinction.

independence condition paralleling (8.8) is then trivially true:

$$\dot{Y}_j \perp\!\!\!\perp \dot{X} \,|\, \dot{X}{=}j, \dot{Z}{=}z \quad (\text{for } j = 0, 1) \tag{8.11}$$

But this condition has not the same interpretation. (8.8) can be interpreted as the requirement that X, conditional on values of Z, is approximately independent of all further circumstances which are fixed by the implicit reference to particular units. (8.11), in contrast, only says that the outcome variable, given $X = j$ and $\dot{Z} = z$, is independent of any other outcome variable for which $\dot{X} \neq j$ and $\dot{Z}{=}z$.

(5) As mentioned, in order to satisfy the independence condition (8.8), there ideally should be a randomized assignment of values of X to the members of Ω (conditional on values of Z). (8.11) does not require any randomization procedure. The important point is, however, that formulating a causal rule like (8.4) does not entail the claim that there are no further variables on which the outcome variable depends. Such variables are simply not taken into account. Consequently, also definitions of causal effects which are derived from (8.4) do not entail anything about further variables on which the outcome variable depends. Consider the causal effect defined in (8.1). This definition compares two generic units, one with $\dot{X}{=}x''$ and the other one with $\dot{X}{=}x'$. Both units have identical values of $\dot{Z}$; but they can differ in all other respects.

6. Omitted causally relevant conditions

If a causal rule does not relate to an artificial random generator, one can almost always think that the rule misses one or more causally relevant conditions. Note that this is true even if the data used to estimate the rule result from a randomized experiment. The point simply is that there probably are causally relevant conditions not explicitly referred to in the rule's formulation. It is therefore not reasonable to require that a causal rule entails the claim that one has taken into account all causally relevant conditions.

Moreover, except when dealing with artificial random generators, already the assumption that one can 'theoretically' refer to a complete set of variables which are causally relevant for an outcome variable seems obscure. The descriptive approach to potential outcomes avoids this assumption and instead requires the conditional independence (8.8). This independence is viewed as a precondition for thinking of a causal effect of X. However, as already mentioned, (8.8) cannot be formulated in a rule-based approach. In a rule-based approach one would need to refer to explicitly defined variables which, in addition to Z, are causally relevant for Y. If one could refer to a list of such variables, say $(U_1, U_2, \ldots)$, one could use the formulation

$$X \perp\!\!\!\perp (U_1, U_2, \ldots) \,|\, Z = z \tag{8.12}$$

However, the formulation is not useful because one cannot define, not even clearly think of, such a list of variables.

Of course, it is often quite possible to think of a particular variable, say U, which is left out in the formulation of a causal rule, but should be taken into account in order to get a better understanding of the causal mechanism. The original model, that was used to derive the causal rule, must then be enlarged by incorporating U; and this also demands to specify U's relationship with the other variables in the model. How to do this depends on the intended use of the model. If the model is intended to represent a randomized experiment, one can assume in the formulation of the enlarged model that $\dot{X} \perp\!\!\!\perp \dot{U} \,|\, \dot{Z} = z$.

However, as I have argued in § 8.2.5, in social research an explanatory model can almost never be formulated as a model representing a randomized experiment. It then depends on the details of the model how to think of $\dot{U}$'s role in the mechanism generating values of the outcome variable. In any case, one would need observations of $\dot{U}$ in order to quantify its causal role.

Chapter 9

Selection and choice

This chapter continues with the discussion of 'causal interpretations' that was begun in the previous chapter. The focus is on 'selection problems'. In the first section, I distinguish between 'sample selection problems' which concern data-generating processes, and selection events in fact-generating processes which have causal effects. This leads to a distinction between problems resulting from selection, and problems resulting from omitted confounders.

In the second section, I consider choice as a specific kind of selection. I distinguish between two uses of choice variables and introduce, correspondingly, explanatory and evaluative choice models.

In the third section, I first note that there are different kinds of choice situations which require different understandings of causal effects. The problem of omitted confounders occurs in only one of these situations where

the choice concerns different ways to reach comparable outcomes. This choice situation also motivates an interest in balanced effect constructions. I finally briefly consider such constructions.

9.1 Different kinds of selection

1. Distinguishing selection problems

I distinguish three situations where Y is a variable of interest, and there is a further variable S which in some sense involves a selection.

a) S is a binary variable, and values of Y can be observed if $S = 1$, and cannot be observed if $S = 0$. This can properly be called a 'sample selection problem' because S only concerns the observability of Y but is not a causally relevant condition for Y. Here it is presupposed that Y has a distribution that exists independently of S.

b) S is a binary variable, and $S = 1$ is a necessary precondition for Y to have a distribution. For example, being employed ($S = 1$) is a necessary precondition for receiving a wage (Y).

c) S can take two or more different values, and it is assumed that the distribution of Y causally depends on the value of S. For example, values of S represent school types, and Y is a measure of educational attainment.

It is obvious that there is a selection problem in the first case (a). It is not clear, however, in which sense there might be a selection problem in cases (b) and (c). In the following, I begin with a very brief consideration of (a) and then focus on (b). The discussion is continued in the next section where also (c) will be considered.

2. Selection in data-generating processes

Sample selection problems can be conceptualized in two ways. First, Y and S are statistical variables, say

$$(Y, S) : \Omega \longrightarrow \mathcal{Y} \times \{0, 1\}$$

One knows the conditional distribution $\mathrm{P}[Y|S=1]$, but is interested in the unconditional distribution $\mathrm{P}[Y]$. As an example, one can think that S is a response indicator in a survey: $S=1$ if a sampled unit provides a value of Y, and $S=0$ otherwise.

Another framework uses random variables, say $\dot{Y}$ and $\dot{S}$. $\dot{S}$ is again a binary variables and records whether a value of $\dot{Y}$ can be observed. So it is assumed that one knows the conditional distribution $\Pr[\dot{Y}|\dot{S}=1]$, but is interested in the unconditional distribution $\Pr[\dot{Y}]$.[1]

If $\dot{S}$ is independent of $\dot{Y}$, one can use $\Pr[\dot{Y}|\dot{S}=1]$ to estimate $\Pr[\dot{Y}]$. Problems occur if, and because, $\dot{S}$ and $\dot{Y}$ are correlated. This suggests to find another variable, say $\ddot{V}$ (often consisting of several components), such that

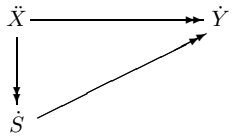
$$\dot{S} \perp\!\!\!\perp \dot{Y} \mid \ddot{V} = v$$

is approximately true. This would allow one to use $\Pr[\dot{Y}|\ddot{V}=v, \dot{S}=1]$ to estimate $\Pr[\dot{Y}|\ddot{V}=v]$.[2]

3. Selection as a necessary precondition

I now consider the case where the selection variable represents a necessary precondition for values of the variable of interest. As an example, I take $\dot{Y}$ to represent the outcome of a university education, and $\dot{S}=1$ if a person is admitted to begin with university studies, and $\dot{S}=0$ otherwise. It is further assumed that $\dot{Y}$ depends on a variable $\ddot{X}$ representing the level of academic performance the person has reached just before the selection variable gets a particular value. A functional model could then be graphically depicted as follows:

Model 9.1.1



The picture is possibly misleading, however, because it suggests that $\dot{Y}$ has a distribution even if $\dot{S}=0$. But, obviously, a process generating a value of $\dot{Y}$ can only take place if $\dot{S}=1$. This entails that there are two rules:

$$x \longrightarrow \Pr(\dot{Y}=y|\ddot{X}=x, \dot{S}=0) = 0 \tag{9.1}$$

and

$$x \longrightarrow \Pr(\dot{Y}=y|\ddot{X}=x, \dot{S}=1) \tag{9.2}$$

Nevertheless, both $\ddot{X}$ and $\dot{S}$ are causally relevant for $\dot{Y}$. Referring to expectations of $\dot{Y}$, $\dot{S}$ can be viewed as an event variable having the effect

[1] Of course, the interest could also concern distributions of $\dot{Y}$ (or Y in the first framework) which depend on further covariates.

[2] This approach is often followed when the selection results from unit nonresponse in sample surveys; see Rohwer (2011).

$\mathrm{E}(\dot{Y}|\ddot{X}=x, \dot{S}=1)$ (see § 8.1.4). Effects of $\ddot{X}$ can be defined, of course, only conditional on $\dot{S}=1$:

$$\mathrm{E}(\dot{Y}|\ddot{X}=x'', \dot{S}=1) - \mathrm{E}(\dot{Y}|\ddot{X}=x', \dot{S}=1) \tag{9.3}$$

Note that in this model there can be no interaction of $\ddot{X}$ and $\dot{S}$ w.r.t. $\dot{Y}$.

4. Counterfactual and modal questions

Although $\dot{S}=1$ is a necessary precondition for values of $\dot{Y}$, one can ask hypothetical questions. Two forms of such questions must be distinguished:

a) *Counterfactual questions* presuppose that $\dot{S}$ already has taken a particular value. For example: Given $\ddot{X}=x$ and $\dot{S}=0$, what value of $\dot{Y}$ might be expected if $\dot{S}$ had taken the value 1 instead of 0? (The complementary question obviously has a trivial answer.)

b) *Modal questions* presuppose a situation where $\dot{S}$ has not already taken a particular value, and a process that might generate a value of $\dot{Y}$ has not yet started.

I will not discuss the counterfactual questions. To answer the modal questions, one can use the rules (9.1) and (9.2), respectively. The selection variable $\dot{S}$ then only serves to distinguish the two modal questions and does not provide any further information. (This will be further discussed in the next section.)

5. Is there a sample selection problem?

The rule (9.1) can be established by referring to the institutional framework without reference to sampled data. Numerically specified versions of the rule (9.2) must be estimated from sampled data. However, as suggested by the rule's formulation, one can simply use the data for those students who actually began a university education ($\dot{S}=1$).

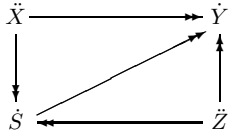
Note that selection on $\dot{S}=1$ has nothing to do with a sample selection problem. The condition $\dot{S}=1$ simply determines the scope of the rule to be estimated. It is possible, of course, that the available data are selective; for example, response rates can depend on $\dot{Y}$. There is then a sample selection problem that leads to biased estimates of $\Pr(\dot{Y}=y|\ddot{X}=x, \dot{S}=1)$. But this bias is not due to conditioning on $\dot{S}=1$, instead, its definition would presuppose this condition.

6. Consideration of omitted confounders

There is, however, another problem that must be considered: the omission of possibly relevant confounding variables. This requires to refer to an enlarged model that explicitly contains a further confounding variable. I

use the following

Model 9.1.2



where $\ddot{Z}$ is a further variable on which $\dot{S}$ depends, e.g., an indicator of a person's educational aspiration. Moreover, it is assumed that also $\dot{Y}$ depends on $\ddot{Z}$ entailing that $\ddot{Z}$ is a confounding variable.

There is no problem if data are available for both $\ddot{X}$ and $\ddot{Z}$. But now assume that data on $\ddot{Z}$ are not available so that one can only estimate a reduced version of the model; in terms of expectations:

$$\begin{aligned}&\mathrm{E}(\dot{Y}|\ddot{X}=x, \dot{S}=1) = \\ &\quad \sum\nolimits_z \mathrm{E}(\dot{Y}|\ddot{X}=x, \dot{Z}=z, \dot{S}=1)\Pr(\dot{Z}=z|\ddot{X}=x, \dot{S}=1)\end{aligned} \tag{9.4}$$

where $\dot{Z}$ is used instead of $\ddot{Z}$ in order to allow thinking of conditional distributions. This shows that effects of $\ddot{X}$, as defined in (9.3), do not have a balanced formulation in the reduced model. Of course, this is the general problem resulting from the omission of confounding variables.

Here it is important, however, that the problem does not result from conditioning on $\dot{S}$. It is true that $\dot{S}$ is a collider, entailing that conditioning on $\dot{S}$ changes the correlation between values of $\ddot{X}$ and $\ddot{Z}$ in a reference set (sample). But this is not relevant here because effect definitions are conditional on $\dot{S}=1$; in the example, they only concern students who actually begin with university studies. Problems resulting from the omission of confounding variables should therefore be clearly distinguished from 'selection problems'.

7. Illustration with a numerical example

To illustrate the argument, I use the fictitious data in Table 9.1.1. Based on these data, one can estimate the rule

$$(x, z) \longrightarrow \mathrm{E}(\dot{Y}|\ddot{X}=x, \dot{Z}=z, \dot{S}=1) \tag{9.5}$$

and finds the values

x	z	$\mathrm{E}(\dot{Y}\|\ddot{X}=x, \dot{Z}=z, \dot{S}=1)$
0	0	0.5
0	1	0.7
1	0	0.8
1	1	0.9

(9.6)

Table 9.1.1 Fictitious data for the example.

x	z	s	$y=0$	$y=1$	cases
0	0	0	–	–	800
0	0	1	100	100	200
0	1	0	–	–	600
0	1	1	120	280	400
1	0	0	–	–	400
1	0	1	120	480	600
1	1	0	–	–	200
1	1	1	80	720	800

Conditioning on $\dot{S}=1$ changes the relationship between $\ddot{X}$ and $\dot{Z}$:

x	$\Pr(\dot{Z}=1\|\ddot{X}=x,\dot{S}=0)$	$\Pr(\dot{Z}=1\|\ddot{X}=x,\dot{S}=1)$
0	0.429	0.667
1	0.333	0.571

(9.7)

This is not relevant, however, because the rule (9.5) only applies to situations where $\dot{S}=1$.

Now assume that values of $\dot{Z}$ are not available. The data then lead to the following estimate of an effect of $\ddot{X}$:

$$\mathrm{E}(\dot{Y}|\ddot{X}=1,\dot{S}=1)-\mathrm{E}(\dot{Y}|\ddot{X}=0,\dot{S}=1)=0.857-0.633=0.224 \quad (9.8)$$

This effect is not balanced. As can be seen from (9.7), the distribution of the omitted variable $\dot{Z}$ is different for $\ddot{X}=0$ and $\ddot{X}=1$. This means that the effect cannot be attributed solely to a difference in $\ddot{X}$. Of course, without observations on $\dot{Z}$ one cannot assess the contribution of this variable.

8. Selection and omitted confounders

In order to stress that 'selection problems' and problems resulting from unobserved confounders should be clearly distinguished, I briefly consider a further example: How does the risk of divorce depends on the women's level of education? Researchers often found that education has a positive impact on the risk of divorce, but there also are other findings. In a recent study, Bernardi and Martinez-Pastor (2011) discuss the hypothesis that the observed relationships between education and risk of divorce might result, at least in part, from 'selection effects'. However, so formulated, the hypothesis can easily create confusion.

To see this, one can use Model 9.1.2 with the following interpretation. $\dot{S}=1$ if a woman is married, and $\dot{S}=0$ otherwise. $\dot{Y}$ is an indicator

variable for becoming divorced;[3] $\ddot{X}$ records the level of education, and $\ddot{Z}$ is an unobserved confounder. Obviously, without being married there can be no risk of divorce. Consequently, $\dot{S}=1$ is also a necessary precondition for the variables', $\ddot{X}$ and $\ddot{Z}$, having an impact on the risk of divorce.

Of course, one can assume that distributions of values of $\ddot{X}$ and $\ddot{Z}$ exist for married and unmarried women in some specified population.[4] So one can think that marriage changes these distributions and their correlation, and consider this as a 'selection effect'. But this selection effect cannot change relationships between these variables and the risk of divorce; simply because these relationships are only defined conditional on $\dot{S}=1$.

However, an important part of the research question concerns the observation of historically changing relationships between women's education and the risk of divorce. One has then to consider at least two periods, say

$$(\ddot{X}_1,\dot{Z}_1,\dot{S}_1,\dot{Y}_1) \longrightarrow (\ddot{X}_2,\dot{Z}_2,\dot{S}_2,\dot{Y}_2) \quad (9.9)$$

$\dot{Z}_t$ represents the distribution of the unobserved confounder in period t. This allows one to think of the observed relationships in the following way:

$$\mathrm{E}(\dot{Y}_t|\ddot{X}_t=x,\dot{S}_t=1)= \quad (9.10)$$
$$\sum\nolimits_z \mathrm{E}(\dot{Y}_t|\ddot{X}_t=x,\dot{Z}_t=z,\dot{S}_t=1)\Pr(\dot{Z}_t=z|\ddot{X}_t=x,\dot{S}_t=1)$$

So it is quite possible that differences between the observed relationships can be due to both

a) changes in the conditional expectation $\mathrm{E}(\dot{Y}_t|\ddot{X}_t=x,\dot{Z}_t=z,\dot{S}_t=1)$ which, presumably, has a causal interpretation, and

b) changes in the conditional distributions of the unobserved confounder, $\Pr[\dot{Z}_t|\ddot{X}_t=x,\dot{S}_t=1]$.

Most probably, they are due to both kinds of changes so that one would like to learn about the quantitative relevance of unobserved confounders. But, of course, this would require to observe the confounder.

The question remains whether one can think of (b) as a 'selection effect'. Obviously not in the static sense referred to above. One would need to consider the function

$$(x,z) \longrightarrow \Pr(\dot{S}_t=1|\ddot{X}_t=x,\dot{Z}_t=z) \quad (9.11)$$

Differences in these functions would describe a historically changing selection into marriage. However, the changes referred to in (b) do not only

[3] Bernardi and Martinez-Pastor use a duration model, but this is not important for the present conceptual discussion.

[4] To make this precise, one would need statistical variables instead of the modal variables, $\ddot{X}$ and $\ddot{Z}$.

result from a change in the function (9.11). As seen from

$$\Pr(\dot{Z}_t=z|\ddot{X}_t=x,\dot{S}_t=1) = \frac{\Pr(\dot{S}_t=1|\ddot{X}_t=x,\dot{Z}_t=z)\Pr(\dot{Z}_t=z|\ddot{X}_t=x)}{\Pr(\dot{S}_t=1|\ddot{X}_t=x)} \tag{9.12}$$

they can also result from a change in the conditional distributions of the unobserved confounder.

Consequently, without observing the confounder (whose supposed existence motivates the discussion) one cannot draw any clear conclusions. On the other hand, if one could observe $\dot{Z}_t$, one could immediately consider the relationship

$$(x,z) \longrightarrow \mathrm{E}(\dot{Y}_t|\ddot{X}_t=x,\dot{Z}_t=z,\dot{S}_t=1) \tag{9.13}$$

and recognize that the risk of divorce not only depends on the women's level of education, but also on another identifiable variable, $\dot{Z}$.

9.2 Choice-based selection

I now consider selection variables which get their values from decisions of individual or collective actors. Such variables will be called 'choice variables'.

1. A notion of choice variables

As I will use the term, a *choice variable*, say C, has the following features:

a) The domain of C is a set of $m \geq 2$ alternatives, numerically represented by $\mathcal{C} = \{1,\ldots,m\}$.

b) Referring to a choice variable entails that there is an individual or collective agent who can choose, or already has chosen, a particular value of the variable. The agent associated with a choice variable C will be denoted by $A[C]$.

c) It is presupposed that the agent has the power to select one of the alternatives. In other words, $\mathcal{C}$ must only contain states which can be realized by the agent.

d) The agent is assumed to have considered the alternatives before one of them is actually chosen. I do not assume that the agent is 'rational' in any particular sense.

Following this definition, choice is a specific kind of selection, namely a selection which is reflexively generated by an actor. This is meant by the expression 'choice-based selection'.[5]

2. Two contexts for using choice variables

Choice variables can be used in two different contexts. In one context, one conceives of a choice variable as a kind of event variable. $C=c$ then means that the agent, $A[C]$, has chosen the alternative $c \in \mathcal{C}$. In addition, $C=0$ means that such an event has not yet occurred.

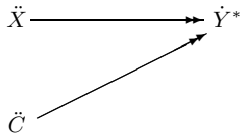
In this understanding, choice variables can be used in explanatory models, both as explanatory variables and as dependent variables. A model that attempts to explain choices, considered as events, will be called an *explanatory choice model*. Such a model can be depicted as

$$\ddot{X} \longrightarrow\!\!\!\!\!\rightarrow \dot{C}$$

where $\dot{C}$ is the choice variable, and $\ddot{X}$ denotes the explanatory variable (possibly consisting of several components). For example, one can think that the model is intended to represent the choice of a child's school type; $A[\dot{C}]$ refers to the child's parents, and $\ddot{X}$ denotes the parents' educational level. Note that this understanding entails that the parents have the power to choose a school type for their child.

In a quite different context, choice variables serve to consider possible consequences of hypothetically chosen alternatives. To hypothetically give a choice variable a value is obviously different from an actual choice, it simply means to consider that alternative. Thus, in this understanding, a choice variable is not an event variable; and in a functional model it can only be used as an exogenous variable. As an example, think again of the parents' choice of a school type for their child. One can consider the following model:

Model 9.2.1



The model relates to a generic $A[\ddot{C}]$, the parents who are assumed to consider possible values of $\ddot{C}$ (school types). As before, $\ddot{X}$ denotes the parents' educational level. $\dot{Y}^*$ is used to assess possible outcomes: the child's educational success that can be expected if, given $\ddot{X}$, the parents would choose $\ddot{C}=c$. The corresponding function is

$$(x,c) \longrightarrow \mathrm{E}(\dot{Y}^*|\ddot{X}=x,\ddot{C}=c) \tag{9.14}$$

[5] So it has nothing to do with 'choice-based sampling', that is, sampling designs which use a stratification w.r.t. realized choices; see Scott and Wild (1989).

A model of this kind will be called an *evaluative choice model*. Its aim is not to predict realized choices. Instead, it is intended to serve thinking about modal questions (in the sense introduced in § 9.1.4).[6] Consequently, also $\dot{Y}^*$ cannot be understood as representing realized outcomes, and must be distinguished from an outcome variable in an explanatory model.

There obviously are similarities between an evaluative choice model and a treatment model (as this term was introduced in § 8.2.2). But also note the different tasks. A treatment model serves to formulate a generic rule about causal effects of treatments (= events which can be deliberately generated). An evaluative choice model serves an agent to consider the consequences of available alternatives. While randomization might be used for a treatment model, randomization w.r.t. the choice variable would contradict the idea of a choice.

3. Primary and secondary actors of choice models

Remember the distinction between primary and secondary actors introduced in Section 8.2. The distinction can easily be applied to explanatory choice models. These models are concerned with choices made by primary actors. The secondary actors, in contrast, are those who construct and use these models for one reason or another.

Now consider an evaluative choice model. Since the model serves an agent to think about available alternatives and possible consequences of choosing one of them, the agent is a secondary actor w.r.t. to the model. On the other hand, the model is concerned with the agent's choice, and so the agent can also be considered as a primary actor. However, an evaluative model has not the task to predict the agent's choice. 'To choose' and 'to predict the outcome of a choice' are obviously different activities.

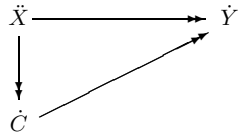
4. Consideration of modal questions

Let me stress the connection between modal questions (as introduced in § 9.1.4) and evaluative choice models. An explanatory choice model cannot be used to consider modal questions (or must be reinterpreted in some way). As an illustration, I refer to Model 9.2.1. This model is intended to show how expectations of an outcome variable, $\dot{Y}^*$, depend on exogenously given values of $\ddot{X}$ and hypothetically chosen values of $\ddot{C}$.

Of course, in order to become useful, one needs a quantification of the function (9.14). Data might come from a sample that relates to the following model where it is assumed that $\dot{Y}$ has the same meaning as $\dot{Y}^*$.

[6] Of course, an evaluative choice model only allows an agent to consider conditional expectations and cannot be used to formally derive a particular decision.

Model 9.2.2



In this model, $\dot{C}$ is an event variable, representing alternatives chosen by the primary agents to which the model relates. Since $\dot{C}$ depends on $\ddot{X}$, part of the model consists in an explanatory choice model.

Only the evaluative model 9.2.1, not the explanatory model 9.2.2, can be used for modal questions. In the explanatory model, values of $\dot{C}$ come into being according to a function, $\ddot{X} \longrightarrow\!\!\!\!\!\rightarrow \dot{C}$, that predicts the choices actually made. This model can therefore not be used for a situation where a choice variable can hypothetically be given different values because a choice event has not yet occurred.

Nevertheless, data corresponding to the explanatory model 9.2.2 can also be used to estimate the function (9.14). Such data would allow one to estimate $\mathrm{E}(\dot{Y}|\ddot{X}=x,\dot{C}=c)$, and then use the rule

$$\text{Estimate } \mathrm{E}(\dot{Y}^*|\ddot{X}=x,\ddot{C}=c) \text{ by } \mathrm{E}(\dot{Y}|\ddot{X}=x,\dot{C}=c) \tag{9.15}$$

As assumed by Model 9.2.2, the data result from 'self selection' in the sense of choices, made by primary actors, which depend on a variable $\ddot{X}$; $\dot{S}$ stochastically depends on $\ddot{X}$.[7] But this does not entail a sample selection problem that might create a bias when using the rule (9.15). As mentioned in § 9.1.6, it is quite possible that Model 9.2.2 misses a relevant confounder and could be replaced by a better model (if the necessary data would be available). But this has nothing to do with the fact that the model contains an event variable that gets its values by choices of primary actors.

5. Modal questions w.r.t. necessary preconditions

I now consider a situation where the choice concerns a necessary precondition for a possible outcome. To illustrate, I continue with the example that was introduced in § 9.1.3: beginning with university studies and consideration of possible outcomes. Details depend on the set-up of the choice situation. I consider two situations.

(1) I begin with a situation where an agent has the power to decide whether a person (the agent herself or someone else) will begin with university studies. There is then a choice variable, $\ddot{C}$, having the domain $\mathcal{C} = \{1, 2\}$; and

[7] Note that this is not the case in the evaluative model 9.2.1. In fact, in that model it is not even possible to imagine a dependence of $\ddot{C}$ on $\ddot{X}$.

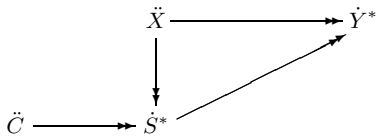
$\ddot{C}=1$ means 'beginning' and $\ddot{C}=2$ means 'not beginning' with university studies. The outcome assumed to be relevant for the choice is the success of the university education; it will be denoted by $\dot{Y}^*$. I further assume that expectations about $\dot{Y}^*$ depend on an exogenous variable, $\ddot{X}$, representing properties of the person who possibly begins university studies that are known in the choice situation.

With this notation, one can again use Model 9.2.1 as an evaluative model and Model 9.2.2 as a corresponding explanatory model. The evaluative model is to be used for the two modal questions. First, what would be the outcome if $\ddot{C}=2$? Obviously, without beginning university studies there could be no successful outcome ($\dot{Y}^*$ is then either undefined or has the value zero)

Now consider the other question, What is the expected value of $\dot{Y}^*$ if $\ddot{C}=1$, given $\ddot{X}=x$? To answer this question, one needs an estimate of $\mathrm{E}(\dot{Y}^*|\ddot{X}=x,\ddot{C}=1)$. This can be derived from the explanatory model 9.2.2 by using the rule (9.15). One only needs $\mathrm{E}(\dot{Y}|\ddot{X}=x,\dot{C}=1)$. Whether also $\mathrm{E}(\dot{Y}|\ddot{X}=x,\dot{C}=2)$ can be given a sensible interpretation is obviously irrelevant.

(2) The argument in (1) presupposes that the choice variables in the two models, $\ddot{C}$ and $\dot{C}$, have the same meaning. I now consider a situation where an agent cannot immediately decide whether to begin, or not begin, with university studies, but only whether to apply for admission. The choice variable, $\ddot{C}$, has again the domain $\mathcal{C}=\{1,2\}$, but now $\ddot{C}=1$ means 'to apply' and $\ddot{C}=2$ means 'not to apply'. An evaluative choice model can now be depicted as follows.

Model 9.2.3



$\ddot{X}$ and $\dot{Y}^*$ (and correspondingly $\dot{Y}$) have the same meaning as before. In addition, there is now the variable $\dot{S}^*$ representing the decision of the admission committee: $\dot{S}^*=1$ if the applicant is admitted, and $\dot{S}^*=0$ otherwise. Both $\dot{Y}^*$ and $\dot{S}^*$ are starred in order to distinguish these variables from corresponding variables in an explanatory model.

The model can be used for thinking about modal questions in two steps. In a first step, the agent, $A[\ddot{C}]$, can consider $\mathrm{E}(\dot{S}^*|\ddot{X}=x,\ddot{C}=1)$. Given information from a corresponding explanatory model (see Model 9.1.1), one can use $\mathrm{E}(\dot{S}|\ddot{X}=x,\dot{C}=1)$ and a suitably modified version of rule (9.15). In a second step, one can think about the final outcome, $\dot{Y}^*$, conditional on $\dot{S}^*=1$. In this step, $\mathrm{E}(\dot{Y}|\ddot{X}=x,\dot{S}=1)$ can be used as an estimate of $\mathrm{E}(\dot{Y}^*|\ddot{X}=x,\dot{S}^*=1)$. Finally, both steps can be combined:

$$\text{Estimate } \mathrm{E}(\dot{Y}^*|\ddot{X}=x,\ddot{C}=1) \text{ by } \mathrm{E}(\dot{Y}|\ddot{X}=x,\dot{S}=1)\,\mathrm{E}(\dot{S}|\ddot{X}=x) \quad (9.16)$$

Notice that the evaluative model 9.2.3 relates to a choice situation where the committee has not yet decided about $A[\ddot{C}]$'s application (simply because $A[\ddot{C}]$ has not yet decided whether to apply). Possibly useful information can therefore only result from estimates of $\mathrm{E}(\dot{S}|\ddot{X}=x)$.

9.3 Causal effects of choice variables

1. Three different choice situations

How to think of causal effects of choice variables depends on the kind of choice situation. I propose to distinguish three situations where $\ddot{C}$ always is a binary choice variable with domain $\mathcal{C}=\{0,1\}$.[8]

a) The choice concerns a necessary precondition for a process generating values of an outcome variable, say $\dot{Y}$, to take place. So the causal effect of $\ddot{C}$ can simply be stated: $\ddot{C}=1$ is a necessary precondition for $\dot{Y}$. In order to quantify the effect, one can consider $\ddot{C}$ as an event variable and use

$$\mathrm{E}(\dot{Y}|\ddot{X}=x,\ddot{C}=1) \quad (9.17)$$

where $\ddot{X}$ represents conditions on which the process generating values of $\dot{Y}$ depends.

b) Associated with the two alternatives are two qualitatively different outcome variables, say $\dot{Y}_0$ and $\dot{Y}_1$. This implies that one has to consider two qualitatively different effects, one effect of $\ddot{C}=0$ and another one of $\ddot{C}=1$. Both should be considered separately as indicated in (a).

c) The choice concerns two different ways to generate values of a single outcome variable, $\dot{Y}$. So one can compare the alternatives w.r.t. expectations of $\dot{Y}$, and use the effect definition

$$\Delta^s(\dot{Y};\ddot{C}[0,1],\ddot{X}{=}x) := \mathrm{E}(\dot{Y}|\ddot{X}{=}x,\ddot{C}{=}1) - \mathrm{E}(\dot{Y}|\ddot{X}{=}x,\ddot{C}{=}0) \quad (9.18)$$

In the following, I only discuss the situation (c), in particular, how to think about unobserved confounders. Note that this problem is of no particular relevance in the situations (a) and (b). Of course, the conditional expectation referred to in (9.17) might well depend on further variables. But the

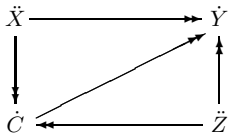
[8] This differs from the convention introduced in § 9.2.1, but eases the notation.

effect of $\ddot{C}$ can always be interpreted as an average effect w.r.t. to distributions of these variables. So this is different from the situation discussed in Section 9.1 where one is interested in effects of $\ddot{X}$ (e.g. education of married women).

2. The problem: omitted confounders

To focus the discussion, I consider

Model 9.3.1



The structure is the same as in Model 9.1.2, but it is now assumed that $\dot{Y}$ has a distribution for both values of $\dot{C}$. As an example, one can think that the choice is between two learning frames, σ_0 and σ_1, where a person can acquire competencies represented by $\dot{Y}$, and it is assumed that the processes generating values of $\dot{Y}$ also depend on two further variables, $\ddot{X}$ and $\ddot{Z}$, having values already fixed in the choice situation.

If data for both variables are available, one can use $\ddot{Z}=z$ as an additional condition in the effect definition (9.18). But if data on $\ddot{Z}$ are not available, this variable is an omitted confounder, and the effect definition (9.18) relates to a reduced model. Substituting $\ddot{Z}$ by a variable $\dot{Z}$ having a distribution, the observable effect is then given by

$$\begin{aligned}\mathrm{E}(\dot{Y}|\dot{C}[0,1],\ddot{X}=x) = \qquad (9.19)\\ \sum\nolimits_z \big[\mathrm{E}(\dot{Y}|\dot{C}=1,\ddot{X}=x,\dot{Z}=z)\Pr(\dot{Z}=z|\dot{C}=1,\ddot{X}=x) -\\ \mathrm{E}(\dot{Y}|\dot{C}=0,\ddot{X}=x,\dot{Z}=z)\Pr(\dot{Z}=z|\dot{C}=0,\ddot{X}=x)\big]\end{aligned}$$

This effect is no longer balanced because

$$\Pr[\dot{Z}|\dot{C}=1,\ddot{X}=x] \neq \Pr[\dot{Z}|\dot{C}=0,\ddot{X}=x]$$

and it is therefore unclear how to think of a causal effect of $\dot{C}$; see the illustration in Figure 9.3.1. This difficulty motivates an interest in balanced effect formulations.

3. The perspective of evaluative choice models

A further motive for an interest in balanced effect formulations comes from evaluative choice models where an agent is interested in potential effects of the available alternatives. Consider the evaluative choice model that corresponds to Model 9.3.1:

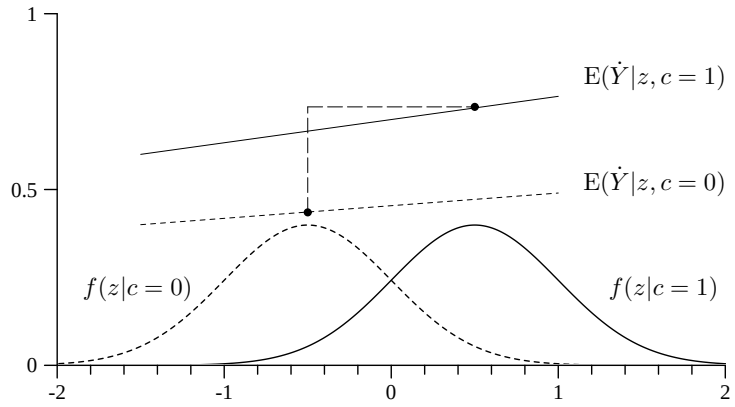
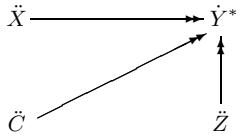


Figure 9.3.1 Illustration of an unbalanced effect (values of $\dot{Z}$ on the abscissa).

Model 9.3.2



The agent, $A[\ddot{C}]$, knows the value of $\ddot{X}$, say x^*. Assuming that only the rule (9.19) is available, expectations are evaluated by comparing

$$\Pr(\dot{Y}^*|\ddot{C}=0,\ddot{X}=x^*) \quad \text{and} \quad \Pr(\dot{Y}^*|\ddot{C}=1,\ddot{X}=x^*)$$

As shown by (9.19), both are averages w.r.t. different distributions of $\dot{Z}$. However, the agent can assume that there is a particular value of this variable, say z^*, that is identical for both alternatives in the given choice situation. Even if this value is not known, it would probably preferable to use a balanced effect formulation.[9]

4. Constructions of balanced effects

So one should ask, Can balanced effect formulations be constructed? There are three approaches.

a) Randomization w.r.t. all possible confounders. As I have argued in the previous chapter, this approach is problematic in social research

[9] In this respect, evaluative choice models are similar to treatment models which also motivate an interest in balanced effect formulations.

because it would change the processes that one aims to investigate.

b) One uses a fixed distribution of the confounding variables. For example, instead of (9.19) one can consider

$$\sum\nolimits_z \big[\mathrm{E}(\dot{Y}|\dot{C}=1,\ddot{X}=x,\dot{Z}=z)\Pr(\dot{Z}=z) - \mathrm{E}(\dot{Y}|\dot{C}=0,\ddot{X}=x,\dot{Z}=z)\Pr(\dot{Z}=z)\big] \tag{9.20}$$

where $\Pr[\dot{Z}]$ is an arbitrarily defined distribution (e.g., the distribution of a corresponding statistical variable in a sample). Of course, this method can only be used with observed confounders; one does not get effects that are balanced w.r.t. unobserved confounders.

c) One uses specific kinds of parametric models which allow one to construct balanced effects. This will be further discussed in the next paragraph.

5. Parametric assumptions about confounders

I use Heckman's probit selection model (Heckman 1979) to illustrate the parametric approach to the construction of balanced effects.[10] The basic idea is to assume that omitted confounders can be implicitly taken into account by a joint parametric distribution for the variables $\dot{Y}$ and $\dot{C}$.

To explain the argument, I start form Model 9.3.1. If $\ddot{Z}$ is observed, one can begin with a linear model

$$\dot{Y} = g_x(x) + g_z(z) + \dot{C}\gamma + \epsilon' \tag{9.21}$$

where g_x and g_z are deterministic functions of values of $\ddot{X}$ and $\ddot{Z}$, respectively, and ϵ' is a residual random variable. Assuming that the distribution of ϵ' is independent of $\dot{C}$, one can interpret γ as a balanced effect. – If $\ddot{Z}$ is not observed, one can consider the reduced model that is based on assuming, instead of $\ddot{Z}$, a variable $\dot{Z}$ with some unknown but exogenously given distribution. Instead of (9.21), one gets the reduced model

$$\dot{Y} = g_x(x) + \dot{C}\gamma + \epsilon \tag{9.22}$$

where $\epsilon := g_z(\dot{Z}) + \epsilon'$. Since $\dot{C}$ depends on $\dot{Z}$, the distribution of ϵ depends on $\dot{C}$, and the effect of $\dot{C}$ is not balanced w.r.t. ϵ. But from (9.22) one can derive

$$\mathrm{E}(\dot{Y}|x,c) = g_x(x) + c\gamma + \mathrm{E}(\epsilon|x,c)$$

[10] Here I discuss this model as a proposal for coping with unobserved confounders ('endogeneity bias'). As proposed by Heckman, the model is also used for 'sample selection problems' in the sense defined in § 9.1.1, that is, in situations where observations are only available if $\dot{C}=1$.

This shows that, in order to estimate γ, one would like to know values of $\mathrm{E}(\epsilon|x,c)$. If observed, they could be used as values of a further variable in a regression model for $\dot{Y}$. Of course, these values cannot be observed; but given enough parametric assumptions, they can be constructed. This is the leading idea of Heckman's approach.

There are two steps. In the first step one assumes a model for $\dot{C}$. This is done by employing a latent variable, η', as follows:

$$\dot{C} = I[\eta' > h_x(x) + h_z(z)] \tag{9.23}$$

where h_x and h_z are deterministic functions of values of $\ddot{X}$ and $\ddot{Z}$, respectively. Again, one can define $\eta := -h_z(z) + \eta'$, and rewrite (9.23) as $\dot{C} = I[\eta > h_x(x)]$.

In the second step, the distributional assumptions come into play. These concern ϵ', η', and $\dot{Z}$:

a) $\epsilon' \sim \mathcal{N}(0, \sigma^2_{\epsilon'})$

b) $\eta' \sim \mathcal{N}(0,1)$

c) $g_z(\dot{Z}) \sim \mathcal{N}(\mu_{g_z}, \sigma^2_{g_z})$

d) $h_z(\dot{Z}) \sim \mathcal{N}(\mu_{h_z}, \sigma^2_{h_z})$

e) ϵ' and $g_z(\dot{Z})$ are independent.

f) η' and $h_z(\dot{Z})$ are independent.

These assumptions entail: $\epsilon \sim \mathcal{N}(\mu_{g_z}, \sigma^2_\epsilon)$ with $\sigma^2_\epsilon = \sigma^2_{g_z} + \sigma^2_{\epsilon'}$, and $\eta \sim \mathcal{N}(\mu_{h_z}, \sigma^2_\eta)$ with $\sigma^2_\eta = \sigma^2_{h_z} + 1$. Moreover, if g_z and h_z are linear functions, the joint distribution of ϵ and η is bivariate normal with a correlation ρ.

From these assumptions and their implications, one can finally derive a method to construct values of $\mathrm{E}(\epsilon|x,c)$:

$$\mathrm{E}(\epsilon|x,c{=}1) = \mathrm{E}(\epsilon|x,\eta > h_x(x)) = \rho\,\sigma_\epsilon\,\frac{\phi(h_x(x))}{1-\Phi(h_x(x))}$$

and correspondingly

$$\mathrm{E}(\epsilon|x,c{=}0) = \mathrm{E}(\epsilon|x,\eta \le h_x(x)) = \rho\,\sigma_\epsilon\,\frac{\phi(h_x(x))}{1-\Phi(-h_x(x))}$$

where ϕ and Φ denote, respectively, the density and distribution function of the standard normal distribution. Both can be combined as

$$\lambda(x,c) := c\,\frac{\phi(h_x(x))}{1-\Phi(h_x(x))} + (1-c)\,\frac{\phi(h_x(x))}{1-\Phi(-h_x(x))}$$

It can now be seen how this is an approach to the construction of balanced effects. By defining a new residual variable,

$$\epsilon^* := \epsilon - \rho\,\sigma_\epsilon\,\lambda(x,c)$$

one can rewrite (9.22) as

$$\dot{Y} = g(x) + \dot{C}\gamma + \rho\,\sigma_\epsilon\,\lambda(x,c) + \epsilon^* \tag{9.24}$$

containing a further regressor, $\lambda(x,c)$, and the new residual, ϵ^*. Now one can derive

$$\mathrm{E}(\epsilon^*|x,c) = 0 \quad \text{and} \quad \mathrm{Cov}(\epsilon^*, \dot{C}|x) = 0$$

showing that the effect of $\dot{C}$ is balanced w.r.t. ϵ^*.

6. Confounding and mediating variables

It is obvious that Heckman's approach relies on very particular assumptions about the distributions of unobserved variables which are difficult to justify in applications.[11] The most important assumption concerns the distribution of the unobserved confounder, $\dot{Z}$. If it is not normally distributed, e.g. if it is a binary variable, the argument depicted in the previous paragraph will not work.

A further point is noteworthy. Heckman's probit selection model deals with 'endogeneity bias' resulting from omitted variables which show up in correlations between explanatory variables and the residual variable posited in a regression model. The model cannot, therefore, distinguish between omitted confounders and omitted mediating variables. However, also omitted mediating variables can result in correlations between explanatory and residual variables.

Consider Model 9.3.1. If the arrow from $\ddot{Z}$ to $\dot{C}$ is reversed, $\ddot{Z}$ changes into an endogenous mediating variable, $\dot{Z}$. Now, if $\dot{Z}$ is omitted from a regression model for $\dot{Y}$, $\dot{C}$ is again correlated with its residual. In fact, the argument of the previous paragraph will go through without any formal changes. But it would be difficult to think that the model removes an 'endogeneity bias'.

[11] For critical discussion see Little (1985), Briggs (2004), Bushway, Johnson and Slocum (2007).

References

Agresti, A., Agresti, B. F. (1977). Statistical Analysis of Qualitative Variation. In: K. F. Schuessler (ed.), *Sociological Methodology 1978*, 204–237. San Francisco: Jossey-Bass.

Allison, P. D. (1999). Comparing Logit and Probit Coefficients Across Groups. *Sociological Methods & Research* 28, 186–208.

Angrist, J. D., Pischke, J.-S. (2009). *Mostly Harmless Econometrics*. Princeton: Princeton University Press.

Baron, R. M., Kenny, D. A. (1986). The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. *Journal of Personality and Social Psychology* 51, 1173–82.

Baumert, J., Watermann, R., Schümer, G. (2003). Disparitäten der Bildungsbeteiligung und des Kompetenzerwerbs. *Zeitschrift für Erziehungswissenschaft* 6, 46–72.

Bäumer, T. et al. (2011). Education Processes in Life-course-specific Learning Environments. *Zeitschrift für Erziehungswissenschaft*, Special Issue 14, 87–101.

Bernardi, F., Martinez-Pastor, J.-I. (2011). Female Education and Marriage Dissolution: Is it a Selection Effect? *European Sociological Review* 27, 693–707.

Blossfeld, H.-P. (2009). Causation as a Generative Process. In: H. Engelhardt, H.-P. Kohler, A. Prskawetz (eds.), *Causal Analysis in Population Studies*, 83–109. Berlin: Springer.

Blossfeld, H.-P., Roßbach, H.-G., Maurice, J. von (eds.) (2011). Education as a Lifelong Process. The German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft*, Special Issue 14.

Boudon, R. (1974). *Education, Opportunity, and Social Inequality.* New York: John Wiley.

Breen, R., Jonsson, J. O. (2000). Analyzing Educational Careers: A Multinomial Transition Model. *American Sociological Review* 65, 754–772.

Briggs, D. C. (2004). Causal Inference and the Heckman Model. *Journal of Educational and Behavioral Statistics* 29, 397–420.

Bushway, S., Johnson, B. D., Slocum, L. A. (2007). Is the Magic Still There? The Use of the Heckman Two-Step Correction for Selection Bias in Criminology. *Journal of Quantitative Criminology* 23, 151–178.

Cameron, S. V., Heckman, J. J. (1998). Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males. *Journal of Political Economy* 106, 262–333.

Cramer, J. S. (2007). Robustness of Logit Analysis: Unobserved Heterogeneity and Mis-specified Disturbances. *Oxford Bulletin of Economics and Statistics* 69, 545–555.

Cross, C. B. (1991). Explanation and the Theory of Questions. *Erkenntnis* 34, 237–60.

Ditton, H., Krüsken, J., Schauenberg, M. (2005). Bildungsungleichheit – der Beitrag von Familie und Schule. *Zeitschrift für Erziehungswissenschaft* 8, 285–304.

Erikson, R., Goldthorpe, J. H., Jackson, M., Yaish, M., Cox, D. R. (2005). On Class Differentials in Educational Attainment, *Proceedings of the National Academy of Sciences* 102, 9730–33.

Erikson, R., Jonsson, J. O. (1996). Explaining Class Inequality in Education: The Swedish Test Case. In: R. Erikson, J. O. Jonsson (eds.), *Can Education be Equalized?* 1–63. Boulder: Westview Press.

Erikson, R., Rudolphi, F. (2010). Change in Social Selection to Upper Secondary School – Primary and Secondary Effects in Sweden. *European Sociological Review* 26, 291–305.

Faye, J. (1999). Explanation Explained. *Synthese* 120, 61–75.

Fisher, R. A. (1922). On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London.* Series A, Vol. 222, 309–368.

Gangl, M. (2010). Causal Inference in Sociological Research. *Annual Review of Sociology* 36, 21–47.

Gerring, J. (2004). What is a Case Study and What is it Good for? *American Political Science Review* 98, 341–354.

Goldrick-Rab, S. (2006). Following Their Every Move: An Investigation of Social-Class Differences in College Pathways. *Sociology of Education* 79, 61–79.

Goldthorpe, J. H. (2001). Causation, Statistics, and Sociology. *European Sociological Review* 17, 1–20.

Greenland, S. (2004). An Overview of Methods for Causal Inference from Observational Studies. In: A. Gelman, X.-L. Meng (eds.), *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, 3–13. New York: Wiley.

Hansen, M. N. (1997). Social and Economic Inequality in the Educational Career: Do the Effects of Social Background Characteristics Decline? *European Sociological Review* 13, 305–321.

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica* 47, 153–161.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association* 81, 945–968.

Holm, A., Jaeger, M. M.. (2011). Dealing with Selection Bias in Educational Transition Models: The Bivariate Probit Selection Model. *Research in Social Stratification and Mobility*, doi: 10.1016/j.rssm.2011.02.002.

Jackson, M., Erikson, R., Goldthorpe, J. H., Yaish, M. (2007). Primary and Secondary Effects in Class Differentials in Educational Attainment. *Acta Sociologica* 50, 211–229.

Kloosterman, R., Ruiter, S., Graaf, P. M. de, Kraaykamp, G. (2009). Parental Education, Children's Performance and the Transition to Higher Secondary Education: Trends in Primary and Secondary Effects Over Five Dutch School Cohorts (1965-99). *British Journal of Sociology* 60, 377–398.

LaLonde, R. J.(1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review* 76, 604–620.

Little, R. J. A. (1985). A Note About Models for Selectivity Bias. *Econometrica* 53, 1469–1474.

MacKinnon, D. P. (2008). *Introduction to Statistical Mediation Analysis*. New York: Lawrence Erlbaum.

Mahoney, J. (2001). Beyond Correlational Analysis: Recent Innovations in Theory and Method. *Sociological Forum* 16, 575–593.

Mare, R. D. (1980). Social Background and School Continuation Decisions. *Journal of the American Statistical Association* 75, 295–305.

Mare, R. D. (1981). Change and Stability in Educational Stratification. *American Sociological Review* 46, 72-87.

Mare, R. D. (1993). Educational Stratification on Observed and Unobserved Components of Family Background. In: Y. Shavit and H.-P. Blossfeld (eds.), *Persistent Inequalities. Changing Educational Attainment in Thirteen Countries,* 351–376. Boulder: Westview Press.

Mood, C. (2010). Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It. *European Sociological Review* 26, 67–82.

Morgan, S. L., Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles*. Cambridge: Cambridge University Press.

Nash, R. (2006). Controlling for 'Ability': a Conceptual and Empirical Study of Primary and Secondary Effects. *British Journal of Sociology of Education* 27, 157–172.

Neugebauer, M. (2010). Bildungsungleichheit und Grundschulempfehlung beim Übergang auf das Gymnasium: Eine Dekomposition primärer und sekundärer Herkunftseffekte. *Zeitschrift für Soziologie* 39, 202–214.

Pearl, J. (2001): Direct and Indirect Effects. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, 411–420. San Francisco, CA.

Petersen, M. L., Sinisi, S. E., Van der Laan, M. J. (2006). Estimation of Direct Causal Effects. *Epidemiology* 17, 276–284.

Rohwer, G. (2010). *Models in Statistical Social Research*. London: Routledge.

Rohwer, G. (2011). Uses of Probabilistic Models of Unit Nonresponse. NEPS Working Paper No. 5. Bamberg: NEPS.

Rohwer, G., Pötter, U. (2002). *Methoden sozialwissenschaftlicher Datenkonstruktion.* Weinheim: Juventa.

Rubin, D. B. (2005). Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association* 100, 322–331.

Rubin, D. B., Stuart, E. A., Zanutto, E. L. (2004). A Potential Outcomes View of Value-Added Assessment in Education. *Journal of Educational and Behavioral Statistics* 29, 103–116.

Schindler, S., Reimer, D. (2010). Primäre und sekundäre Effekte der sozialen Herkunft beim Übergang in die Hochschulbildung. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 62, 623–653.

Schindler, S., Lörz, M. (2011). Mechanisms of Social Inequality Development: Primary and Secondary Effects in the Transition to Tertiary Education Between 1976 and 2005. *European Sociological Review*.

Schneider, S. L., Tieben, N. (2011). A Healthy Sorting Machine? Social Inequality in the Transition to Upper Secondary Education in Germany. *Oxford Review of Education* 37,, 139–166.

Scott, A. J., Wild, C. J. (1989). Selection Based on the Response Variable in Logistic Regression. In: C. J. Skinner, D. Holt, T. M. F. Smith (eds.), *Analysis of Complex Surveys*, 191–205. New York: Wiley.

Shavit, Y., Blossfeld, H.-P. (eds.) (1993). *Persistent Inequalities. Changing Educational Attainment in Thirteen Countries.* Boulder: Westview Press.

Tieben, N., Wolbers, M. (2010). Success and Failur in Secondary Education: Socio-economic Background Effects on Secondary School Outcome in the Netherlands, 1927-1998. *British Journal of Sociology of Education* 31, 277–290.

Weinberg, C. R. (1993). Toward a Clearer Definition of Confounding. *American Journal of Epidemiology* 137, 1–8.

Winship, C., Morgan, S. L. (1999). The Estimation of Causal Effects from Observational Data. *Annual Review of Sociology* 25, 659–706.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data.* Cambridge: MIT Press.

Index