Regression Models for Clustered and Interdependent Observations

Ulrich Pötter

Götz Rohwer

Paper prepared for the LWBB-workshop on "Multilevel Analysis of Longitudinal Data," Potsdam, November 28 – 29, 1997.

Author's address: Ruhr-Universität Bochum. Fakultät für Sozialwissenschaft. Universitätsstrasse 150, GB-E1. 44801 Bochum, Germany.

Summary

The paper tries to show that, for explanatory purposes, multilevel random coefficient models do not provide advantages not already available with standard regression models containing interaction terms. An alternative interpretation is based on the idea that these models can be used for scaling a population of groups. Finally, the paper argues that except for a few special cases, the general idea of interdependent individuals creates difficulties for any kind of regression models.

Introduction

A basic argument for using multilevel models refers to the fact that people do not behave independently from each other but are linked together by a variety of social relations. This is certainly true, and as it is an important goal of statistical models in empirical social research to provide insight into dependency relations, it is certainly a good idea to develop models that can incorporate information about social relations. The question is, however, how to construct such models. The current discussion focuses on random coefficient models (e.g., Bryk and Raudenbush 1992, Longford 1993, 1995, Goldstein 1995, Blien et al. 1994). In our discussion, we shall try to raise some doubts about this approach. Section 1 tries to show that, from the point of view of explanatory model construction, multilevel random coefficient models do not provide advantages not already available with standard regression models containing interaction effects. Section 2 considers the idea that these models can be used, alternatively, for "scaling groups", that is, to estimate group-specific characteristics. Section 3 provides some critical remarks with regard to the often expressed opinion that the introduction of group-level variables conflicts with assumptions required for OLS. The final Section 4 deals with the idea of interdependent individuals. It is argued that, except for some very special cases, this fact creates difficulties for any kind of regression models.

1 Random Coefficient Models

Consider two variables, X and Y. Assuming a linear relationship, we can set up an ordinary regression model

$$Y = \alpha + X\beta + \epsilon \tag{1}$$

Given observations $(x_i, y_i; i = 1, \dots, n)$ and using some estimation strategy (e.g., OLS), we find parameter estimates, $\hat{\alpha}$ and $\hat{\beta}$, and corresponding residuals $\hat{\epsilon}_i$. The observations might be the outcome of an experiment or of observing some aspects of social reality. We are primarily interested in the latter situation. We can then imagine that the individual units who generated the data are connected by a set of social relations. The literature explaining multilevel models normally begins with the most simple situation where there is a single equivalence relation inducing a partition of the individual units into a set of groups, say, $\mathcal{G}_1, \dots, \mathcal{G}_m$.

Let us assume that there is an additional variable, Z, characterizing in some way these groups. For the moment, we shall assume that Z is a single metric variable. A situation where we have, instead, a set of dummy variables will be discussed below.

The normal approach to use this information for an elaboration of model (1) would be to include Z as an additional regressor variable. And, of course, it would also be possible to consider interaction effects resulting in a model

$$Y = \alpha_0 + X\beta_x + Z\beta_z + XZ\beta_{xz} + \eta \tag{2}$$

As a direct implication, we get a different model for each value of Z, namely

$$Y = (\alpha_0 + z\beta_z) + X(\beta_x + z\beta_{xz}) + \eta_z \tag{3}$$

Given sufficient data, model (2) can be estimated with OLS or some other estimation method. And it would be possible to partition the variance of the estimated residuals according to the groups induced by Z.[1]

In recent literature on multilevel random coefficient models (MLRCM) it has become popular to use a somewhat different modeling approach. The basic approach (see, e.g., Mason, Wong and Entwisle 1984, De Leeuw and Kreft 1986) does not directly include Z into (1) but follows a two-step procedure. In a first step, the model parameters in (1) are interpreted as random variables. And in a second step Z is used to establish additional regression models for these random variables:

$$\alpha = \alpha_0 + Z\beta_z + \epsilon_\alpha \tag{4}$$

$$\beta = \beta_x + Z\beta_{xz} + \epsilon_\beta \tag{5}$$

Having established these additional models, they can be combined with (1) and the final MLRCM becomes

$$Y = \alpha_0 + X\beta_x + Z\beta_z + XZ\beta_{xz} + (\epsilon_\alpha + X\epsilon_\beta + \epsilon) \tag{6}$$

This model is obviously identical with (2) except for a more complex error term. One might ask, therefore, whether there are any reasons for using (6) instead of (2).

[1] We mention this because some authors (e.g., DiPrete and Forristal 1994, p. 334) seem to believe that calculation of variance components is only possible with random coefficient models.

1. We first note that referring to groups, in the sense of schools, occupations, regions etc., is not essential, neither for (2) nor for (6). As shown by (2), the model does not make any difference between X and Z. Calling Z a "group-level variable" might, in fact, be misleading because what is important is not that Z characterizes some a priori given groups but that Z *induces* a partition of observational units according to its possible values. But this is true for all kinds of regressor variables. Consequently, the distinction between individual-level and group-level variables becomes obscure. For example, the variable "sex" is commonly interpreted as an individual-level variable, but can equally well be viewed as a group-level variable because it simply distinguishes, and characterizes, two groups, men and women. On the other hand, all so-called group-level variables can equally well be interpreted as individual-level variables simply by using these variables to characterize individuals. For example, sex ratio in school classes can be used to characterize their individual members. In fact, a regression model treats all pupil having the same value of this variable alike, regardless of their actual class membership. It is questionable, therefore, whether, in the context of regression models, the multilevel rhetoric has established an important distinction between different types of variables.

2. One argument given by MLRCM proponents for using (6) instead of (2) is that this model can provide additional information about how the effect of X on Y depends on the context. This is most often seen as the main advantage of multilevel models (see, e.g., Bryk and Raudenbush 1992, p. 6; Goldstein 1995, p. 17). Mason et al. (1984, pp. 74-5) give the following formulation:

> "Our fundamental assumption is that the micro values of the response variable in some way depend on context and that the *effects* of the micro determinants *may* vary systematically as a function of context."

In a similar formulation, Blien et al. (1994, p. 270) say:

> "In the analysis of a two level structure we are mainly interested in the variation of the regression coefficients across groups."

Leaving aside the multilevel rhetoric, both statements refer to the question how the effect of X on Y depends on Z. If we interpret this question as referring to the conditional expectation of the outcome variable, both models, (2) and (6), give the same answer. Using $\mathrm{E}(Y|x,z)$ to denote the conditional expectation of Y on values of X and Z, we get

$$\partial \mathrm{E}(Y|x,z)/\partial x = \beta_x + z\beta_{xz} \tag{7}$$

3. A somewhat different version of the argument is that we are not just interested in how the effect of X on Y depends on the context, Z, but we want to estimate a distribution of these varying effects (see, e.g., Goldstein 1995, p. 17). There is, unfortunately, some ambiguity in this word, "distribution", when referring to model parameters. If we finally want *empirical* information, the word "distribution" must refer to the given distribution of Z. Then, by substituting z by Z in (7) we get a distribution of the effects of X on Y as a linear function of Z. And given sufficient data, this distribution can be estimated empirically.

Therefore, this additional argument does also not provide a reason why we should add an error term to (7). (5) makes β a random variable, namely a function of the random variable Z, already without adding ϵ_β. This error term only adds confusion since we need then a conceptually different probability space for interpretation. At the beginning we are only given some probability space for (X,Y,Z). We are then free to assume a model for some aspects of the relationship between these variables. This implies a definition of model parameters. Finally, when using a standard regression model, we get as an implication of the model an additional random variable, the error term. This error term is defined then with respect to the same probability space which is assumed for (X,Y,Z). However, such a view on model construction will not lead to (4) and (5). To make sense of these equations, one would need a conceptually different probability space.[2]

4. Assume that we are not given a single metric variable, Z, but instead, a set of dummy variables indicating group membership, say, $D_1, \ldots, D_m$. The analogue of (2) would then become[3]

$$Y = \alpha_0 + X\beta_x + \sum\nolimits_j D_j\delta_j + \sum\nolimits_j XD_j\delta_{x,j} + \eta \tag{8}$$

and the analogue of (7) would become

$$\partial \mathrm{E}(Y|x)/\partial x = \beta_x + \sum\nolimits_j D_j\delta_{x,j} \tag{9}$$

So far, also the interpretation would be the same. (9) can be interpreted as showing the distribution of the effects of X on Y, now given by the regression parameters for the set of dummy variables. To get a close correspondence between (8) and (2), we may also construct variables D and D_x reflecting the parameters δ_j and $\delta_{x,j}$, respectively. (If an individual unit belongs to $\mathcal{G}_j$, its values for D and D_x would be δ_j and $\delta_{x,j}$, respectively.) Model (8) can then be written as

$$Y = \alpha_0 + X\beta_x + D + XD_x + \eta \tag{10}$$

showing a direct correspondence with (2).

5. Here begins another argument to motivate the MLRCM approach. The argument refers to a situation where the number of groups is fairly large such that we cannot expect all groups represented in our samples. The standard example refers to schools. For instance, Goldstein (1995, p. 17) writes:

> "If we wish to focus not just on these schools [given in the sample], but on a wider 'population' of schools, then we need to regard the chosen schools as giving us information about the characteristics of all the schools in the population. Just as we choose random samples of individuals to provide estimates of population means etc., so a randomly chosen sample of schools can provide information about the characteristics of the population of schools. In particular, such a sample can provide estimates of the variation and covariation between schools in the slope and intercept parameters and will allow us to compare schools with different characteristics."

Of course, if the number of groups, m, becomes fairly large then also the number of parameters in (8) and this might result in estimation problems. However, before dealing with estimation problems, the first question should be why we might be interested in estimating (8) if there is a huge number of different groups.

- If the number of groups is small it makes sense to interpret group membership as an additional factor influencing the outcome variable. A

[2] This does not create any problems from a Bayesian point of view, see, e.g., Lindley and Smith (1972). However, this view is normally not followed in the MLRCM literature.

[3] To ease the comparison of different models we assume that the parameters for the dummy variables are deviations from some mean value, α_0.

necessary condition for this interpretation is that we are able to distinguish and identify the different groups. In particular, this interpretation requires that we can associate with each group a qualitatively different situation. For example, variables like sex, birth cohorts and region can be used in this way. If we think, on the other hand, of typical clusterings into large number of different groups (e.g., households or schools), the conditions for an interpretation as explanatory variables are not fulfilled. For explanatory purposes, the important factor is then not that an individual belongs to one specific group, and not to any of the other groups. Instead, one should try to find additional factors which can be used to characterize the different groups. In general, this will then show that not the original clustering (schools or households) is important but a quite different clustering *induced* by the additional explanatory variables.

Therefore, if the number of groups becomes fairly large there is normally no point in trying to estimate (8) if we are interested in an explanatory model.

- We might be interested, nevertheless, in estimating a distribution of regressor effects across groups. As shown in (10), the theoretical idea of such a distribution can be given an empirical sense regardless of the number of different groups. But in our argument, (10) was derived from (8), it simply summarizes a feature of model (8). The RCM approach goes the other way around. (8) is discarded because of estimation problems and, instead, one begins directly with a suitable parameterization of (10) by assuming D and D_x to be random variables with a known distribution.

However, what can be learned? In the standard linear MLRC models we will get, at best, some information about the variances of D and D_x since the shape of their distributions is assumed to be already known. The resulting information is then that part of the variance of the residuals can be formally attributed to a clustering into groups. While this might be interpreted as providing some information about the "population of groups", it clearly gives no information about individual group characteristics (to be used for explanatory purposes).

- The question remains why we should have any interest in the distribution of regressor effects across groups if we are not able to interpret this distribution in terms of interaction effects. As we have noted, this interpretation depends in a crucial way on the possibility to interpret membership in specific groups as a potential factor for the outcome variable. In fact, the RCM approach explicitly destroys such an interpretation by assuming that group membership is random. This implies that we explicitly forget about the identity of groups.[4] And, consequently, membership in specific groups cannot be used for explanatory purposes.

2 Scaling Groups

One might wonder why some authors are so worried by the task of estimating a distribution of regression parameters in a situation where the number of groups is so large that membership in specific groups cannot, by itself, be used for explanatory purposes. The following remark by Goldstein (1995, p. 17) might throw some light on this question:

> "An important class of situations arises when we wish primarily to have information about each individual school in a sample, but where we have a large number of schools so that (2.2) [our equation (8) without the interaction terms] would involve estimating a very large number of parameters. Furthermore, some schools may have rather small numbers of students and application of (2.2) would result in imprecise estimates. In such cases, if we regard the schools as members of a population and then use our population estimates of the mean and between-school variation, *we can utilize this information to obtain more precise estimates for each individual school* [our emphasis]."

If we understand correctly, the leading idea is to use a distribution of group-specific parameters to compare individual groups. A typical example (see Bryk and Raudenbush 1992, p. 13) would be a regression of some achievement measure on socioeconomic status with groups defined as schools; interpreting the group-specific intercepts and slopes as indicating "effectiveness" and "equity", respectively, one might be interested in comparing individuals schools with respect to these characteristics. The goal of model construction is then no longer directed towards an explanation of an outcome variable but towards "scaling groups".

We shall try to use this idea as an alternative motivation for using RCM. For simplicity, we consider a model like (8) without interaction terms. The model provides group-specific intercepts, $\alpha_j = \alpha_0 + \delta_j$, which can be used to compare the groups. Now assume that the number of

[4] Longford (1993, p. 11) makes this point very clear when he writes: "In our setup the focus is on a *sample* of clusters; the clusters can be thought of as anonymously labelled units, in the same way as can the elementary observations."

groups becomes very large and (8) is no longer estimable. The RCM approach proposes to substitute the set of fixed effects, δ_j, by a single random variable, ϵ_α, resulting in a model

$$Y = \alpha_0 + X\beta_x + \epsilon_\alpha + \eta \tag{11}$$

Instead of the fixed intercepts, $\alpha_j = \alpha_0 + \delta_j$, this model defines groups by realizations of the random variable $\alpha = \alpha_0 + \epsilon_\alpha$. Of course, (11) is not estimable without additional assumptions. Standard assumptions would be

$$\epsilon_\alpha \sim \mathcal{N}(0, \sigma_\alpha^2), \quad \eta \sim \mathcal{N}(0, \sigma_\eta^2), \quad \mathrm{cov}(\epsilon_\alpha, \eta) = 0 \tag{12}$$

While this approach will not lead to a better explanation of the dependent variable (as we have tried to show in the previous section), the question now is what it implies for the task of scaling the groups. We have two remarks.

1. The first one refers to the assumptions given in (12). The idea of "scaling groups" implies that we are interested in the distribution of group-specific model parameters. In our simple example, this is the distribution of group-specific intercepts. If we intend to give this distribution an empirical interpretation we should think of a variable D, as defined in the previous section. Its distribution is the result of group-specific intercepts and the distribution of individuals across groups. Both are empirical facts, to be estimated, and there is, consequently, no reason why we should expect a normal distribution as implied by (12). (If we compare this with a situation where groups are induced by a metric variable, Z, then the assumption that ϵ_α is normally distributed would imply that also Z is normally distributed. But, of course, the distribution of Z should be viewed as empirically given.)

2. The second remark is concerned with the implications of the RCM approach for scaling the groups. How can we find estimates of group-specific realizations of ϵ_α? Goldstein (1995, p. 24) proposes to consider the expectation of ϵ_α, given the data and all parameter estimates from (11). It is not clear, however, how to follow this proposal. The problem already begins with the question how to speak of different groups when using model (11). If we follow its formulation verbatim, groups are *defined* by different realizations of ϵ_α and do not exist prior to the data generating process behind the model's random variables. However, the whole idea of scaling groups is, of course, based on the assumption that groups do exist in some sense as a real grouping of individuals in social reality.

a) In order to illuminate this incoherence, let us first assume that we are able to consider a *specific* group, $\mathcal{G}_j$. We can then assume group-specific data given by variables X_j and Y_j and, conditional on model (11), our data become values of

$$\tilde{Y}_j = Y_j - \alpha_0 - X_j\beta_x$$

The assumption that we are focusing on a specific group immediately implies that there is a specific realized value of ϵ_α, say $\epsilon_{\alpha,j}$. Then, given $\tilde{Y}_j$, the only remaining random variation is induced by η and we have $\tilde{Y}_j = \epsilon_{\alpha,j} + \eta$. Consequently, we can view $\epsilon_{\alpha,j}$ as the expectation of $\tilde{Y}_j$ with respect to the distribution of η, that is

$$\epsilon_{\alpha,j} = \mathrm{E}_\eta(\tilde{Y}_j) \tag{13}$$

establishing a direct correspondence to an approach that uses dummy variables for group membership. In fact, if we had used (8) (without interaction effects) instead of (11), we would get $\mathrm{E}_\eta(\tilde{Y}_j) = \delta_j$.

b) A different point of view is taken by most authors in the RCM literature (see, e.g., Goldstein 1995, p. 24). In order to understand this appraoch, we consider the following data generating process behind (11):

$$\begin{aligned} \epsilon_\alpha &\rightsquigarrow X_1, \eta_1, \ldots, X_k, \eta_k \\ &\rightsquigarrow \underline{\tilde{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_k),\ \tilde{Y}_i = Y_i - \alpha_0 - X_i\beta_x = \epsilon_\alpha + \eta_i \end{aligned}$$

In a first step, one randomly selects a value of ϵ_α, meaning one member of the "population of groups"; and in a second step, one randomly selects k individuals out of this group. The idea is, that we have observations for $\underline{\tilde{Y}}$, but do not know which value was drawn for ϵ_α. So we want to predict this value, given observations for $\underline{\tilde{Y}}$. The proposal, most often considered in the RCM literature, is to use the conditional expectation

$$\mathcal{P}(\epsilon_\alpha) = \mathrm{E}(\epsilon_\alpha \,|\, \underline{\tilde{Y}} = \underline{\tilde{y}}) \tag{14}$$

to predict this value ($\underline{\tilde{y}}$ denotes the observed values for the vector $\underline{\tilde{Y}}$).

In order to find a solution, Goldstein (l.c.) proposes to consider the regression

$$(\epsilon_\alpha, \ldots, \epsilon_\alpha)' = \underline{\tilde{Y}}\gamma + (\eta_1, \ldots, \eta_k)' \tag{15}$$

Notice that ϵ_α, on the left-hand side, is not indexed by i in order to express the idea that there is a single realization for all individuals in the same group. As a consequence, the covariance matrix becomes

$$\left(\mathrm{cov}(\underline{\tilde{Y}}, \underline{\tilde{Y}})\right)_{ii'} = \begin{cases} \sigma_\alpha^2 + \sigma_\eta^2 & \text{if } \; i = i' \\ \sigma_\alpha^2 & \text{otherwise} \end{cases} \tag{16}$$

This then implies[5]

$$\begin{aligned} \mathrm{E}(\epsilon_\alpha \,|\, \underline{\tilde{Y}} = \underline{\tilde{y}}) &= \mathrm{cov}(\epsilon_\alpha, \underline{\tilde{Y}})\, \mathrm{cov}(\underline{\tilde{Y}}, \underline{\tilde{Y}})^{-1}\, \underline{\tilde{y}} \qquad (17) \\ &= \frac{k\sigma_\alpha^2}{k\sigma_\alpha^2 + \sigma_\eta^2} \sum_{i=1}^{k} \tilde{y}_i / k \end{aligned}$$

and this formula is proposed by Goldstein (l.c.) to predict values of ϵ_α. Compared with (13), the difference is the "shrinkage factor"

$$0 < \gamma = k\sigma_\alpha^2 / (k\sigma_\alpha^2 + \sigma_\eta^2) < 1 \tag{18}$$

Are there reasons to use this covariance matrix for scaling groups? As a procedure for ranking, it posesses some formal optimality properties to recommend itself.[6] However, Golstein's argument, namely that (13) provides "more precise estimates for each individual school" is only tenable if "precision" refers to a statistical measure that averages possible discrepancies over all groups. So it is quite irrelevant for the judgement of a particular group.

3 Regressors and Errors

While scaling groups might be interesting in its own right, it does not contribute to an explanation of the outcome variable. We therefore return to the question in Section 1 whether we should use (6) instead of (2) to estimate an explanatory model. One often repeated argument is that OLS estimation of (2) would be wrong because some of the "usual assumptions" of OLS are not satisfied if observational units are clustered into groups. One argument refers to the "assumption" of stochastic independence between regressor variables and error term. Specifically, formulation (6) seems to suggest that the compound error term, $\epsilon_\alpha + X\epsilon_\beta + \epsilon$, is correlated with regressor X. Another more general argument refers to "dependencies" among observational units when they are clustered into groups. In this section we shortly comment on the first argument, the second one will be dealt with in Section 4.

Let us begin with the simple model (1). The question whether X and ϵ are correlated depends on our understanding of ϵ. There are two possibilities. One is to assume that ϵ does exist, in some way, in social reality *independent of our model formulation*. This "platonistic" view allows to formulate assumptions about the error term, but has two serious implications. First, since the error term is by definition unobservable, such assumptions can never be tested. Second, if not only X and Y, but also ϵ do exist as some facts in reality, there is not the slightest reason why we should expect a linear relation between these variables. And again, since the error variables are not observable, there is no chance to check this assumption.

An alternative view can avoid these implications by recognizing what we are really doing when constructing statistical models. We are given some observable variables like X and Y and we then construct some model for the conditional distribution of Y given X. Many different possibilities are available. Regression models are a special type where we assume a relation between the expectation of Y and values of X. This then *creates* the error variable, namely $\epsilon = Y - \mathrm{E}(Y|X)$. Error variables are therefore derived from assuming a specific type of model.

Now let us assume model (1). The error variable is then defined by $\epsilon = Y - \alpha - X\beta$. However, this definition is incomplete until we either fix the values of the parameters, α and β, or fix the correlation between ϵ and X at zero (or some other value). In order to learn something about the conditional distribution of Y on X, the first approach would be senseless. Consequently, to give ϵ a precise meaning we need to fix sufficient details about the joint distribution of ϵ and X. This then provides an opportunity to estimate the model parameters.

Many different methods are available in order to estimate them. If we *demand* independence between ϵ and regressors, we would normally use OLS or GLS. Both *imply* that estimated residuals and observed regressor variables are uncorrelated. And this would also be true if these methods were used to estimate (2) or (6).[7]

[5] The formula for inverting $\mathrm{cov}(\underline{\tilde{Y}}, \underline{\tilde{Y}})$ can be found in Rao 1973, p. 67.

[6] See Searle et al. (1992, p. 268). Unfortunately they do not provide a proof.

[7] Whether, in fact, OLS or GLS estimation of (2) or (6) results in heteroskedastic residuals should be viewed as an empirical question. The empirical results could then be used to re-evaluate the model specification.

4 Interdependent Individuals

A theoretically more interesting argument in the recent MLRCM literature refers to the fact that, in social reality, observational units do not behave independent from each other. For example, Blien et al. (1994, p. 268-9) say:

> "The assumption of stochastic independence may not be fulfilled when the individuals are clustered within groups, such as classes or schools in education or regions in the labor market. For the individuals that belong to the same cluster, the errors may be correlated."

This is then taken as a further argument against using OLS to estimate (2) or (6). In our view, the argument is somewhat more complicated and finally leads to a principal limitation for the application of regression models, including MLRCM, to model the behavior of interdependent individuals.

1. We first note that the formulation "correlated errors" has not, just from the beginning, a clear meaning. The term "correlation" refers to variables, so we need at least two variables to speak of a correlation. However, consider model (1). There is only a single error variable, ϵ, and, so far, it simply doesn't make sense to speak of correlated errors.

2. This fact is somewhat obscured by the widespread habit to write already the theoretical model in terms of observational units. Following this practice, (1) would become

$$Y_i = \alpha + X_i\beta + \epsilon_i \quad \text{for} \quad i = 1, \ldots, n \tag{19}$$

with index i referring to observational units. Such a formulation obviously provides the opportunity to think of correlated errors, namely $\text{cov}(\epsilon_i, \epsilon_{i'}) \neq 0$. However, to make sense of the formulation we need to understand the meaning of the indexed variables.

3. Let us first try to understand the meaning of (X, Y) in the simple model (1). If we view this as a random variable we are referring to a chance situation, say $\mathcal{S}$, that can create values for (X, Y). We can also think of a data generating process (DGP) that creates these values. In social science applications this DGP takes place in social reality and should not be confused with "sampling data", meaning the process through which we get information about the results of the DGP.

Model (1) is then understandable as a model for the chance situation $\mathcal{S}$. And if we focus on the conditional distribution of Y on X, we can imagine that each value of X defines a somewhat different chance situation $\mathcal{S}_x$. In any case, the model does not refer to any specific individual. If we want, nonetheless, think in terms of individuals then the model refers to the "population" of individuals who might be in the chance situation $\mathcal{S}$, or $\mathcal{S}_x$, depending on the interpretation of the model.

4. Now, why should we want to index our variables with i? One possibility is to understand this as a convenient way to represent data, that is, the result from observing a set of outcomes from the chance situation $\mathcal{S}$. The set of indexed variables, (X_i, Y_i), refers then by definition to the same chance situation, $\mathcal{S}$, and we should assume that all these indexed variables have the same distribution as (X, Y). If the sampling procedure is purely random we should also assume that the variables (X_i, Y_i) are independent, again by definition. Of course, indexing the variables with i provides an opportunity to think about correlations. However, this question refers then to the sampling procedure and is quite independent of the question whether the individuals being in the chance situation $\mathcal{S}$ are in some way related. In any case, as long as we are referring to a single chance situation, $\mathcal{S}$, there is no possibility to reflect the notion of interdependent individuals in terms of correlated variables.

5. What is required to think of two (or more) variables being correlated? There are two essential points. The first one is that the variables are defined on the same sample space. This then allows to think of a joint distribution and, consequently, of a correlation. Think of two variables, X and Y. If they are defined on the same sample space, say Ω, we can imagine a sequence of realization

$$(X(\omega_i), Y(\omega_i)) \quad i = 1, 2, 3, \ldots$$

This then creates the notion of a joint distribution and it would become possible to describe one aspect of this joint distribution in terms of a correlation. If we interpret ω_i as representing randomly drawn individuals, we can assume a joint distribution for any set of variables which are defined simultaneously for all individuals (being in the chance situation behind Ω). However, it does not make sense to think of a correlation between, say, $X(\omega_i)$ and $X(\omega_{i'})$.

Now, we can also view a sample as containing observations for n individuals simultaneously. Assuming a single variable, say X, the sample can be written

$$\{X(\omega_1), \dots, X(\omega_n)\}$$

Alternatively, we can also think in terms of a separate variable for each randomly drawn individual, say

$$\{X_1(\omega_1), \dots, X_n(\omega_n)\}$$

However, these variables are no longer defined on the same sample space, Ω, but on the product space, $\Omega \times \cdots \times \Omega$. While this does not, in general, exclude the possibility to think of a joint distribution for these variables, it requires that we are able to view $\omega = (\omega_1, \dots, \omega_n)$ as an ordered n-tupel. Only then can we think of $(X_1, \dots, X_n)$ as an n-dimensional variable with a joint distribution. This, then, is the second essential requirement for assumptions about correlated variables.

A sampling unit $\omega = (\omega_1, \dots, \omega_n)$ will be called *structured* if the ordinal numbers $i = 1, \dots, n$ can be given some substantive meaning. A simple example would be couples, represented by (ω_1, ω_2) where ω_1 always refers to the man and ω_2 always refers to the women (or the other way around). Then it clearly makes sense to assume that two variables, X_1 defined for men and X_2 defined for women, might be correlated. In this example the structure on the sampling units is given by a structural relationship in couples. Another example would be panel data. The structured observational unit is then $\omega = (\omega_1, \dots, \omega_T)$ where each ω_t refers to the same individual observed for a sequence of time points, $t = 1, \dots, T$. Again, it would make sense to assume that variables $X_1, \dots, X_T$ might be correlated.

However, a standard random sample is, by definition, not structured. The individual observations in the sample are "anonymously labelled" and their order has no meaning. Consequently, although we can represent the data for each individual in the sample by a separate variable, it makes no sense to assume that these variables might be correlated.

6. The conclusion is that there are severe limitations if one tries to capture the notion of related and interdependent individuals in terms of correlated variables representing in some way the individuals. The basic precondition is that we can define a structured observational unit. In general, this seems not possible if we directly refer to individuals. We need, instead, some notion of "social position."

We are sceptical, therefore, whether it would be possible to capture the idea of some relationship between individuals being in the same group (e.g., school) can be captured in terms of correlated variables. This would require that we can view the group as a structured unit. While this might be possible by explicitly referring to the social relations that constitute the group, simply referring to the fact that individuals belong to the same group is not sufficient.

Of course, given groups $\mathcal{G}_j$ (directly, or induced by some variable, Z) we can define for each individual a position by referring to the group the individual belongs to. Each group is then a separate chance situation and we can set up a separate model for each group, say $Y_j = \alpha + X_j\beta + \epsilon_j$, and speculate about correlations between ϵ_j and $\epsilon_{j'}$. But this possibility is clearly not meant when people think of "correlated errors" because individuals are clustered into groups.[8]

7. There remains the question how to reflect the fact that individuals are connected by social relations in the set up of statistical models. Standard regression models seem to provide only two possibilities.

a) One can try to represent characteristics of each individual's social position by regressor variables. This simply means that we add variables that provide information about how an individual is related to other people. But these variables have then the same logical status as all other regressor variables. They cannot be used to give some meaning to the notion of "correlated individuals".

b) Another possibility is to set up regression models for structured observational units. However, this then requires that we are able to define a set of related social positions (independent of the sampling procedure). We have already mentioned couples as a simple example. It should be obvious, however, that this approach is only applicable for very simple relations and not suitable for approaching the general notion of social relations.

References

Blien, U., Wiedenbeck, M., Arminger, G. [1994]. Reconciling Macro and Micro Perspectives by Multilevel Models: An Application to Regional Wage Differences. In: I. Borg, P. P. Mohler (eds.), in: Trends and Perspectives in Empirical Social Research, pp. 266 – 282. Berlin: de Gruyter 1994, 266 - 282

[8] In fact, when estimating MLRCM it is normally assumed that errors across groups are not correlated, see, e.g., Blien et al. 1994, p. 270.

Bryk, A. S., Raudenbush, S. W. [1992]. Hierarchical Linear Models: Applications and Data Analysis Methods. Newbury Park: Sage.

DeLeeuw, J., Kreft, I. [1986]. Random Coefficient Models for Multilevel Analysis. Journal of Educational Statistics 11, 57 – 85.

DiPrete, T. A., Forristal, J. D. [1994]. Multilevel Models: Methods and Substance. Annual Review of Sociology 20, 331 – 357.

Goldstein, H. [1995]. Multilevel Statistical Models (2nd ed). London: Edward Arnold.

Lindley, D. V., Smith, A. F. M. [1972]. Bayes Estimates for the Linear Model. (With Discussion.) Journal of the Royal Statistical Society B 34, 1 – 41.

Longford, N. T. [1993]. Random Coefficient Models. Oxford: Clarendon.

Longford, N. T. [1995]. Random Coefficient Models. In: G. Arminger, C. C. Clogg, M. E. Sobel (eds.), Handbook of Statistical Modeling for the Social and Behavioral Sciences, pp. 519 – 578. New York: Plenum Press.

Mason, W. M., Wong, G. Y., Entwisle, B. [1984]. Contextual Analysis Through the Multilevel Linear Model. In: S. Leinhardt (ed.), Sociological Methodology 1983-84, pp. 72 – 103. San Francisco: Jossey-Bass.

Rao, C. R. [1973]. Linear Statistical Inference and Its Applications. New York: Wiley.

Searle, S. R., Casella, G., McCulloch, C. E. [1992]. Variance Components. New York: Wiley.