1

Introduction to Multilevel Analysis

Jan de Leeuw[1] and Erik Meijer[2]

[1] Department of Statistics, University of California at Los Angeles
[2] University of Groningen, Faculty of Economics and RAND Corporation

1.1 History

A common assumption in much of classical statistics is that observations are independently and identically distributed (or i.i.d.). In regression analysis, using the linear model, we cannot insist on identical distributions, because observations differ in expected value, but we generally continue to insist on independence. In fact, we continue to assume that the stochastic parts of the model, i.e., the *errors* or *disturbance terms*, are still i.i.d.

In educational statistics, and in various areas of quantitative sociology, researchers early on began looking for statistical techniques that could incorporate both information about individuals and information about groups to which these individuals belonged. They realized that one of the most challenging aspects of their discipline was to integrate *micro* and *macro* information into a single model. In particular, in the applications educational statisticians had in mind, students are nested in classes, and classes are nested in schools. And perhaps schools are nested in districts, and so on. We have predictors for variables of all these *levels*, and the challenge is to combine all these predictors into an appropriate statistical analysis, more specifically a regression analysis.

Previously, these problems had been approached by either *aggregating* individual-level variables to the group level or *disaggregating* group-level variables to the individual level. It was clear that both these two strategies were unpleasantly ad hoc and could introduce serious biases. Trying to integrate the results of such analyses, for instance by using group-level variables in individual-level regressions, was known as *contextual analysis* [9] or *ecological regression* [42]. It resulted in much discussion about *cross-level inference* and the possibility, or even the unavoidability, of committing an *ecological fallacy* [104].

J. de Leeuw, E. Meijer (eds.), *Handbook of Multilevel Analysis*,

In *school effectiveness research*, which became popular in the 1970s following the epochal studies of Coleman et al. [22] and Jencks et al. [62], educational researchers realized early on that taking group structure into account could result in dependencies between the individual observations. Economists and biostatisticians involved in agriculture and breeding had realized this earlier and had designed *variance and covariance component models* for the Analysis of Variance. But in school effectiveness research a somewhat different paradigm developed, which looked at dependencies in a more specific way. The emphasis was on regression analysis and on data of two levels, let's say students and schools. Performing a regression analysis for each school separately was not satisfactory, because often samples within schools were small and regression coefficients were unstable. Also, these separate analyses ignored the fact that all the schools were part of the same school system and that, consequently, it was natural to suppose the regression coefficients would be similar. This similarity should be used, in some way or another, to improve stability of the regression coefficients by what became known as *borrowing strength*. Finally, in large scale studies there were thousands of schools and long lists of regression coefficients did not provide enough data reduction to be useful.

On the other hand, requiring the regression coefficients in all schools to be the same was generally seen as much too restrictive, because there were many reasons why regressions within schools could be different. In some schools, test scores were relatively important, while in others, socio-economic status was a much more dominant predictor. Schools clearly differed in both average and variance of school success. Of course, requiring regression coefficients to be constant did provide a large amount of data reduction, and a small sampling variance, but the feeling was that the resulting regression coefficients were biased and not meaningful.

Thus, some intermediate form of analysis was needed, which did not result in a single set of regression coefficients, but which also did not compute regression coefficients separately for each school. This led naturally to the idea of random coefficient models, but it left open the problem of combining predictors of different levels into a single technique. In the early 1980s, Burstein and others came up with the idea of using the first-stage regression coefficients from the separate within-school regressions as dependent variables in a second-stage regression on school-level predictors. But in this second stage, the standard regression models that assumed independent observations could no longer be used, mainly because they resulted in inefficient estimates of the regression coefficients and biased estimates of their standard errors. Clearly, the first-stage regression coefficients could have widely different standard errors, because predictors could have very different distributions in different schools. The size of the school, as well as the covariance of the predictors within schools, determined the dispersions of the within-school regression coefficients. Typical

of this stage in educational research are Langbein [71], Burstein et al. [15], and Burstein [14]. Attempts were made to estimate the second-stage regression coefficients by weighted least squares techniques, or to adjust in some other way for the bias in the standard errors [11, 50, 118]. These attempts were not entirely successful, because at the time the statistical aspects of these two-stage techniques were somewhat baffling. A more extensive historical overview of contextual analysis and Burstein's *slopes-as-outcomes* research is in de Leeuw and Kreft [28] and Kreft and de Leeuw [67].

It became clear, in the mid-1980s, that the models the educational researchers were looking for had already been around for quite some time in other areas of statistics. Under different names, to be sure, and usually in a slightly different form. They were known either as *mixed linear models* [51] or, in a Bayesian context, as *hierarchical linear models* [72]. The realization that the problems of contextual analysis could be imbedded in this classical linear model framework gave rise to what we now call *multilevel analysis*. Thus, multilevel analysis can be defined as the marriage of contextual analysis and traditional statistical mixed model theory.

In rapid succession the basic articles by Mason et al. [81], Aitkin and Longford [2], de Leeuw and Kreft [28], Goldstein [44], and Raudenbush and Bryk [100] appeared. All these articles were subsequently transformed into successful textbooks [46, 67, 76, 101]. The two major research groups in educational statistics led, respectively, by Goldstein and by Raudenbush produced and maintained major software packages [97, 102]. These textbooks and software packages, together with subsequent textbooks, such as Snijders and Bosker [111] and Hox [59], solidified the definition and demarcation of the field of multilevel analysis.

1.2 Application Areas

We have seen that multilevel analysis, at least as we have defined it, started in the mid-1980s in educational measurement and sociology. But it became clear quite rapidly that once you have discovered ways to deal with hierarchical data structures, you see them everywhere. The notion of individuals, or any other type of objects, that are naturally nested in groups, with membership in the same group leading to a possible correlation between the individuals, turned out to be very compelling in many disciplines. It generalizes the notion of intraclass correlation to a regression context. Moreover, the notion of regressing regression coefficients, or using slopes-as-outcomes, is an appealing way to code interactions and to introduce a particular structure for the dependencies within groups.

Survey Data

Many surveys are not simple random samples from a relatively homogeneous population, but are obtained from nested sampling in heterogeneous subgroups. Larger units (e.g., states) are drawn first; within these larger units, smaller units (e.g., counties) are drawn next; and so forth. Large surveys typically contain multiple levels of nesting. Sometimes, all units from a certain level are included, as with stratification. See, e.g., Muthén and Satorra [84] for some examples of the complicated sampling schemes used in survey design. The reason for such a complicated nesting structure of surveys is, of course, that it is assumed that the units are different in some respect. It is then natural to model the heterogeneity between groups through multilevel models. See, e.g., Skinner et al. [109] for a book-length discussion of many aspects of the analysis of survey data.

Repeated Measures

In *repeated measures models* (including *growth study models*) we have measurements on a number of individuals that are replicated at a number of fixed time points. Usually there is only a single outcome variable, but the generalization to multivariate outcomes is fairly straightforward. In addition, it is not necessary that all individuals be measured at the same time points. There can be missing data, or each individual can be measured at different time points. The number of books and articles on the analysis of repeated measures is rapidly approaching infinity, but in the context of multilevel analysis, the key publications are Strenio et al. [116] and Jennrich and Schluchter [63]. Chapter 7 of this volume discusses models for longitudinal data. For an extensive treatment of these longitudinal models in the more general context of mixed linear models, we refer to Verbeke and Molenberghs [122].

A different type of "repeated measures" is obtained with *conjoint choice* or *stated preference* data. With such data, subjects are asked to choose between several hypothetical alternatives, e.g., different products or different modes of transport, defined by a description of their alternatives. When subjects are given more than one choice task, a multilevel structure is induced by the repeated choices of the same individual. The corresponding models for such data are usually more straightforward multilevel models than in the case of longitudinal data, where problems such as dynamic dependence, causing non-interchangeability of the observations, and attrition (selective dropout of the sample) often have to be faced. See, e.g., Rouwendal and Meijer [105] for a multilevel logistic regression (or mixed logit) analysis of stated preference data. Similar data are common in experimental psychology, where multiple experiments are performed with the same subjects.

Twin Studies

In school-based attainment studies we often deal with a fairly small number of rather large groups. But the opposite can also occur, either by the nature of the problem or by design. We can decide to use only a small number of students from each class. Or, in repeated measures studies, we can only have two measurements per individual (a "before" and "after", for instance, with a treatment in between). Another "small groups" example is the twin study, in which group size is typically two. See Chapter 5 for a discussion of this type of data.

Meta-Analysis

Data, including historical data, are now much more accessible than in the past. Many data sets are online or are included in some way or another with published research. This makes it attractive to use previous data sets studying the same scientific problem to get larger sample sizes and perhaps a larger population to generalize to. Such (quantitative) analysis of data or results from multiple previous studies is called *meta-analysis*. In Raudenbush and Bryk [99], multilevel techniques specifically adapted to meta-analysis were proposed. Compare also Raudenbush and Bryk [101, Chapter 7].

Multivariate Data

There is a clever way, used by Goldstein [46, Chapter 6], to fit general multivariate data into the multilevel framework. If we have n observations on m variables, we can think of these m observations as nested in n groups with m group members each. This amounts to thinking of the $n \times m$ data matrix as a long vector with nm elements and then building the model with the usual regression components and a suitable specification for the dispersion of the within-group disturbances. It is quite easy to incorporate missing data into this framework, because having data missing simply means having fewer observations in some of the groups. On the other hand, in standard multilevel models, parameters such as regression coefficients are the same for different observations within the same group, whereas in multivariate analysis, this is rarely the case. Thus, writing the latter as a multilevel model requires some care.

1.3 Chapter Outline

In this first chapter of the Handbook we follow the general outline of de Leeuw and Kreft [29]. After this introduction, we first discuss the *statistical models* used in multilevel analysis, then we discuss the *loss functions* used to measure

badness-of-fit, then the *techniques* used to minimize the loss functions, and, finally, the *computer programs* written for these techniques. By using these various steps in the development of multilevel statistical methods, it is easy to discuss the contributions of various authors. It can be used, for instance, to show that the most influential techniques in the field carefully discuss (and implement) all these sequential steps in the framework. After a section on sampling weights, we give an empirical illustration, in which much of the theory discussed in this chapter will be applied. We close with a few final remarks and appendixes that discuss notation and other useful technical background.

1.4 Models

A statistical model is a functional relationship between random variables. The observed data are supposed to be a realization of these random variables, or of a measurable function of these random variables. In most cases, random variables are only partly specified because we merely assert that their distribution belongs to some parametric family. In that case, the model is also only partly specified, and one of the standard statistical chores is to estimate the values of the unknown parameters.

In this section we discuss the multilevel model in the linear case in which there are, at least initially, only two levels. Nonlinear and multivariate generalizations will be discussed in later chapters of this handbook. We also relate it to variance components and mixed models, which, as we have mentioned above, have been around much longer.

Notation is explained in detail in Appendix 1.A. Our main conventions are to underline random variables and to write vectors and matrices in boldface.

1.4.1 Mixed Models

The *mixed linear model* or MLM is written as

$$\underline{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\underline{\boldsymbol{\delta}} + \underline{\boldsymbol{\epsilon}}, \tag{1.1}$$

with $\boldsymbol{X}[n, r]$, $\boldsymbol{Z}[n, p]$, and

$$\begin{pmatrix} \underline{\boldsymbol{\epsilon}} \\ \underline{\boldsymbol{\delta}} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\emptyset} \\ \boldsymbol{\emptyset} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\emptyset} \\ \boldsymbol{\emptyset} & \boldsymbol{\Omega} \end{pmatrix} \right).$$

To simplify the notation, we suppose throughout this chapter that both $\boldsymbol{X}$ and $\boldsymbol{Z}$ have full column rank.

The regression part of the model has a component with fixed regression coefficients and a component with random regression coefficients. Clearly,

$$\underline{\boldsymbol{y}} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{V}),$$

with

$$\boldsymbol{V} \triangleq \boldsymbol{Z}\boldsymbol{\Omega}\boldsymbol{Z}' + \boldsymbol{\Sigma}. \tag{1.2}$$

This illustrates the consequences of making regression coefficients random. We see that the effects of the predictors in $\boldsymbol{Z}$ are shifted from the expected values to the dispersions of the normal distribution. We also see that MLM is a linear regression model with a very specific dispersion structure for the residuals. The form of the dispersion matrix for the residuals in (1.2) is somewhat reminiscent of the common factor analysis model [63], and this similarity can be used in extending multilevel models to covariance structure and latent variable models (see Chapter 12).

It is convenient to parametrize both dispersion matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ using vectors of parameters $\boldsymbol{\sigma}$ and $\boldsymbol{\xi}$. From now on we actually assume that $\boldsymbol{\Sigma}$ is *scalar*, i.e., $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$. A scalar dispersion matrix means we assume the disturbances $\underline{\boldsymbol{\epsilon}}$ are *homoskedastic*. This guarantees that if there are no random effects, i.e., if $\underline{\boldsymbol{\delta}}$ is zero almost everywhere, then we recover the classical linear model. We also parametrize $\boldsymbol{\Omega}$ as a *linear structure*, i.e., a linear combination of known matrices $\boldsymbol{C}_g$. Thus,

$$\boldsymbol{\Omega} = \xi_1\boldsymbol{C}_1 + \cdots + \xi_G\boldsymbol{C}_G = \sum_{g=1}^{G}\xi_g\boldsymbol{C}_g, \tag{1.3}$$

and, consequently, $\boldsymbol{V}$ also has linear structure

$$\boldsymbol{V} = \xi_1\boldsymbol{Z}\boldsymbol{C}_1\boldsymbol{Z}' + \cdots + \xi_G\boldsymbol{Z}\boldsymbol{C}_G\boldsymbol{Z}' + \sigma^2\boldsymbol{I} = \sum_{g=1}^{G}\xi_g\boldsymbol{Z}\boldsymbol{C}_g\boldsymbol{Z}' + \sigma^2\boldsymbol{I}.$$

The leading example is obtained when $\boldsymbol{\Omega} = (\omega_{kl})$ is completely free, apart from symmetry requirements. Then

$$\boldsymbol{\Omega} = \omega_{11}(\boldsymbol{e}_1\boldsymbol{e}_1') + \omega_{21}(\boldsymbol{e}_2\boldsymbol{e}_1' + \boldsymbol{e}_1\boldsymbol{e}_2') + \cdots + \omega_{pp}(\boldsymbol{e}_p\boldsymbol{e}_p'),$$

with $\boldsymbol{e}_k$ the k-th unit vector, i.e., the k-th column of $\boldsymbol{I}$, $\{\xi_1, \ldots, \xi_G\} = \{\omega_{11}, \omega_{21}, \ldots, \omega_{pp}\}$, and $\{\boldsymbol{C}_1, \ldots, \boldsymbol{C}_G\} = \{\boldsymbol{e}_1\boldsymbol{e}_1', \boldsymbol{e}_2\boldsymbol{e}_1' + \boldsymbol{e}_1\boldsymbol{e}_2', \ldots, \boldsymbol{e}_p\boldsymbol{e}_p'\}$. Another typical example is a restricted version of this where ω_{kl} is a given constant (such as 0) for some values of (k, l). These two examples cover the vast majority of specifications used in multilevel analysis.

In some cases it is useful to write models in scalar notation. Scalar notation is, in a sense, more constructive because it is closer to actual implementation on a computer. Also, it is useful for those who do not speak matrix algebra. In this notation, (1.1) becomes, for example,

$$\underline{y}_i = \sum_{q=1}^{r} x_{iq}\beta_q + \sum_{s=1}^{p} z_{is}\underline{\delta}_s + \underline{\epsilon}_i,$$

or

$$\underline{y}_i = x_{i1}\beta_1 + \cdots + x_{ir}\beta_r + z_{i1}\underline{\delta}_1 + \cdots + z_{ip}\underline{\delta}_p + \underline{\epsilon}_i.$$

A *two-level MLM*, which explicitly takes the group structure into account, is given by

$$\underline{\boldsymbol{y}}_j = \boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\underline{\boldsymbol{\delta}}_j + \underline{\boldsymbol{\epsilon}}_j, \tag{1.4a}$$

with $j = 1, \ldots, m$, and

$$\begin{pmatrix}\underline{\boldsymbol{\epsilon}}_j\\ \underline{\boldsymbol{\delta}}_j\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\emptyset\\ \emptyset\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma}_j & \emptyset\\ \emptyset & \boldsymbol{\Omega}_j\end{pmatrix}\right). \tag{1.4b}$$

and, using $\perp$ for independence,

$$(\underline{\boldsymbol{\epsilon}}_j, \underline{\boldsymbol{\delta}}_j) \perp (\underline{\boldsymbol{\epsilon}}_\ell, \underline{\boldsymbol{\delta}}_\ell) \tag{1.4c}$$

for all $j \neq \ell$.

As before, we assume that $\boldsymbol{\Sigma}_j = \sigma_j^2\boldsymbol{I}$, while, in addition, we assume that $\boldsymbol{\Omega}_j = \boldsymbol{\Omega}$. Thus,

$$\underline{\boldsymbol{y}}_j \sim \mathcal{N}(\boldsymbol{X}_j\boldsymbol{\beta}, \boldsymbol{V}_j),$$

with

$$\boldsymbol{V}_j \triangleq \boldsymbol{Z}_j\boldsymbol{\Omega}\boldsymbol{Z}_j' + \sigma_j^2\boldsymbol{I},$$

and the $\underline{\boldsymbol{y}}_j$ for different j are independent.

Observe that the assumption that the $\boldsymbol{X}_j$ and the $\boldsymbol{Z}_j$ have full column rank can be quite restrictive in this case, because we could be dealing with many small groups (as in Chapter 5).

In most applications of multilevel analysis, it is assumed that all σ_j^2 are the same, so $\sigma_j^2 = \sigma^2$ for all j. This is not always a realistic assumption and, therefore, most of our discussion will use separate variances. This has its drawbacks as well, because, obviously, the number of parameters increases with the number of groups in the sample. Thus, when the sample consists of, say, 1000 schools, we would estimate 1000 variance parameters, which is unattractive. Furthermore, consistent estimation of σ_j^2 requires group sizes to diverge to infinity, and therefore in a practical sense, good estimators of σ_j^2 would require moderate within-group sample sizes (e.g., $n_j = 30$). In applications with many small groups, this is obviously not the case.

We can view $\sigma_j^2 = \sigma^2$ as a no-between-groups variation specification and all σ_j^2 treated as separate parameters as a fixed effects specification. From this, it seems that it would be in the spirit of multilevel analysis to treat σ_j^2 as a random parameter, $\underline{\sigma}_j^2$, and use a specification like

$$\log\underline{\sigma}_j^2 = \boldsymbol{z}_{j,p+1}'\boldsymbol{\gamma}_{p+1} + \underline{\delta}_{j,p+1},$$

with, say, $\underline{\delta}_{j,p+1} \sim \mathcal{N}(0, \omega_{p+1,p+1})$, which may be correlated with the other random terms. Such a specification is uncommon in multilevel analysis, but it

would be particularly straightforward to incorporate in the Bayesian approach to multilevel analysis (Chapter 2). In the Bayesian approach, it is more common to use Gamma or inverse Gamma distributions for variance parameters though, but adaptation of this specification to such distributions is fairly easy.

We will not further discuss specification of $\underline{\sigma_j^2}$ as random parameters in this chapter, and treat σ_j^2 as separate parameters. For a specification with $\sigma_j^2 = \sigma^2$, most expressions are unaltered except for dropping the j subscript. However, there are some instances where the differences are a little bit more pronounced, e.g., in the derivatives of the loglikelihood functions. Then we will indicate how the expressions change. Thus, we cover both specifications.

1.4.2 Random Coefficient Models

The *random coefficient model* or RCM is the model with

$$\begin{aligned}\underline{\boldsymbol{y}} &= \boldsymbol{X}\underline{\boldsymbol{\beta}} + \underline{\boldsymbol{\epsilon}},\\ \underline{\boldsymbol{\beta}} &= \boldsymbol{\beta} + \underline{\boldsymbol{\delta}},\end{aligned}$$

with

$$\begin{pmatrix}\underline{\boldsymbol{\epsilon}}\\ \underline{\boldsymbol{\delta}}\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\boldsymbol{\emptyset}\\ \boldsymbol{\emptyset}\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma} & \boldsymbol{\emptyset}\\ \boldsymbol{\emptyset} & \boldsymbol{\Omega}\end{pmatrix}\right).$$

Obviously, in an RCM we have

$$\underline{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}\underline{\boldsymbol{\delta}} + \underline{\boldsymbol{\epsilon}},$$

which shows that the RCM is an MLM in which $\boldsymbol{Z} = \boldsymbol{X}$.

The RCM in this form is not very useful, because without additional assumptions, it is not identified. We give it in this form here to introduce the notion of random coefficients and to prepare for the multilevel RCM.

The two-level RCM that has been studied most extensively looks like

$$\underline{\boldsymbol{y}}_j = \boldsymbol{X}_j\underline{\boldsymbol{\beta}}_j + \underline{\boldsymbol{\epsilon}}_j, \tag{1.5a}$$

$$\underline{\boldsymbol{\beta}}_j = \boldsymbol{\beta} + \underline{\boldsymbol{\delta}}_j, \tag{1.5b}$$

with the same distributional assumptions as above for the two-level MLM. Observe that the fixed part of $\underline{\boldsymbol{\beta}}_j$ is assumed to be the same for all groups. This is necessary for identification of the model.

In this form the random coefficient model has been discussed in the econometric literature, starting from Swamy [117]. It has also become more popular in statistics as one form of the *varying coefficient model*, although this term is mostly used for models with (partly) systematic or deterministic variation of the coefficients, such as a deterministic function of time or some other explanatory variable [54, 61].

The fact that we are dealing with a two-level model here is perhaps clearer if we use scalar notation. This gives

$$\begin{aligned}\underline{y}_{ij} &= x_{ij1}\underline{\beta}_{j1} + \cdots + x_{ijp}\underline{\beta}_{jp} + \underline{\epsilon}_{ij},\\ \underline{\beta}_{js} &= \beta_{js} + \underline{\delta}_{js}.\end{aligned}$$

An important subclass of the RCM is the *random intercept model* or RIM. It is the same as RCM, except for the fact that we assume that all regression coefficients that are not intercepts have no random component. Thus, all slopes are fixed. For a two-level RIM, we consequently have, with some obvious modifications of the notation,

$$\begin{aligned}\underline{\boldsymbol{y}}_j &= \underline{\mu}_j\mathbf{1}_{n_j} + \boldsymbol{X}_j\boldsymbol{\beta} + \underline{\boldsymbol{\epsilon}}_j,\\ \underline{\mu}_j &= \mu + \underline{\delta}_j.\end{aligned}$$

There is an extensive discussion of RIMs, with many applications, in Longford [76]. The econometric panel data literature also discusses this model extensively; see, e.g., Chamberlain [18], Wooldridge [126, Chapter 10], Arellano [4, Chapter 3], or Hsiao [60, Chapter 3]. Observe that for a RIM,

$$\boldsymbol{V}_j = \omega^2\boldsymbol{E} + \sigma_j^2\boldsymbol{I},$$

where $\boldsymbol{E}$ has all its elements equal to $+1$. This is the well-known intraclass covariance structure, with intraclass correlation

$$\rho_j^2 = \frac{\omega^2}{\omega^2 + \sigma_j^2}\ .$$

1.4.3 Slopes-as-Outcomes Models

We are now getting close to what is usually called multilevel analysis. The *slopes-as-outcomes model* or SOM is the model with

$$\begin{aligned}\underline{\boldsymbol{y}} &= \boldsymbol{X}\underline{\boldsymbol{\beta}} + \underline{\boldsymbol{\epsilon}},\\ \underline{\boldsymbol{\beta}} &= \boldsymbol{Z}\boldsymbol{\gamma} + \underline{\boldsymbol{\delta}},\end{aligned}$$

with $\boldsymbol{X}[n,p]$, $\boldsymbol{Z}[p,r]$, and

$$\begin{pmatrix}\underline{\boldsymbol{\epsilon}}\\ \underline{\boldsymbol{\delta}}\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\emptyset\\ \emptyset\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma} & \emptyset\\ \emptyset & \boldsymbol{\Omega}\end{pmatrix}\right).$$

The characteristic that is unique to this model, compared to others discussed here, is that the random coefficients $\underline{\boldsymbol{\beta}}$ are themselves dependent variables in a second regression equation. Of course, in a SOM we have

$$\underline{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{X}\underline{\boldsymbol{\delta}} + \underline{\boldsymbol{\epsilon}},$$

which shows that the SOM is an MLM in which the fixed regressors are $\boldsymbol{X} = \boldsymbol{X}\boldsymbol{Z}$ and the random regressors are $\boldsymbol{X}$.

The two-level SOM is

$$\underline{\boldsymbol{y}}_j = \boldsymbol{X}_j\underline{\boldsymbol{\beta}}_j + \underline{\boldsymbol{\epsilon}}_j, \tag{1.6a}$$

$$\underline{\boldsymbol{\beta}}_j = \boldsymbol{Z}_j\boldsymbol{\gamma} + \underline{\boldsymbol{\delta}}_j, \tag{1.6b}$$

again with the same distributional assumptions. Here $\boldsymbol{X}_j[n_j, p]$ and $\boldsymbol{Z}_j[p, r]$. It is possible, in principle, to have different numbers of predictors in the different $\boldsymbol{X}_j$, but we will ignore this possibility. The regression equations (1.6b) for the random coefficients imply that differences between the regression coefficients of different groups are partly explained by observed characteristics of the groups. These equations are often of great substantive interest.

By substituting the second-level equations (1.6b) in the first-level equations (1.6a) and by stacking the resulting m equations, we find

$$\underline{\boldsymbol{y}} = \boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{X}\underline{\boldsymbol{\delta}} + \underline{\boldsymbol{\epsilon}},$$

with

$$\boldsymbol{U} \stackrel{\Delta}{=} \begin{pmatrix} \boldsymbol{X}_1\boldsymbol{Z}_1 \\ \vdots \\ \boldsymbol{X}_m\boldsymbol{Z}_m \end{pmatrix} \tag{1.7}$$

and with the remaining terms stacked in the same way, except $\boldsymbol{X}$, which has the direct sum form

$$\boldsymbol{X} = \bigoplus_{j=1}^{m} \boldsymbol{X}_j = \begin{pmatrix} \boldsymbol{X}_1 & & \boldsymbol{\emptyset} \\ & \ddots & \\ \boldsymbol{\emptyset} & & \boldsymbol{X}_m \end{pmatrix}.$$

Again, this shows that the two-level SOM is just an MLM with some special structure. We analyze this structure in more detail below.

In the first place, the dispersion matrix of $\underline{\boldsymbol{y}}$ has block-diagonal or direct-sum structure:

$$\underline{\boldsymbol{y}} \sim \mathcal{N}\Big(\boldsymbol{U}\boldsymbol{\gamma}, \bigoplus_{j=1}^{m} \boldsymbol{V}_j\Big),$$

with

$$\boldsymbol{V}_j \stackrel{\Delta}{=} \boldsymbol{X}_j\boldsymbol{\Omega}_j\boldsymbol{X}_j' + \sigma_j^2\boldsymbol{I}.$$

Second, the design matrix $\boldsymbol{U}$ in the fixed part has the structure (1.7). In fact, there usually is even more structure than that. In the two-level SOM, we often have

$$\boldsymbol{Z}_j = \bigoplus_{s=1}^{p} \boldsymbol{z}_j'; \tag{1.8}$$

i.e., $\boldsymbol{Z}_j$ is the direct sum of p row vectors, all equal to a vector $\boldsymbol{z}_j'$ with q elements. The vector $\boldsymbol{z}_j$ describes group j in terms of q second-level variables. More elaborately,

$$\boldsymbol{Z}_j = \begin{pmatrix} \boldsymbol{z}_j' & \emptyset & \emptyset & \cdots & \emptyset \\ \emptyset & \boldsymbol{z}_j' & \emptyset & \cdots & \emptyset \\ \emptyset & \emptyset & \boldsymbol{z}_j' & \cdots & \emptyset \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \emptyset & \emptyset & \emptyset & \cdots & \boldsymbol{z}_j' \end{pmatrix}.$$

This is easily generalized to direct sums of different vectors, even if they have different numbers of elements. It follows that, if we partition $\boldsymbol{\gamma}$ accordingly into p subvectors of length q, we have

$$E(\underline{\beta}_{js}) = \boldsymbol{z}_{js}'\boldsymbol{\gamma}_s.$$

Also

$$\boldsymbol{U}_j = \boldsymbol{X}_j\boldsymbol{Z}_j = \begin{bmatrix} \boldsymbol{x}_{j1}\boldsymbol{z}_{j1}' & \boldsymbol{x}_{j2}\boldsymbol{z}_{j2}' & \cdots & \boldsymbol{x}_{jp}\boldsymbol{z}_{jp}' \end{bmatrix},$$

where $\boldsymbol{x}_{js}$ is the s-th column of $\boldsymbol{X}_j$. Thus, $\boldsymbol{U}$ is a block-matrix, consisting of m by p blocks, and each block is of rank 1. Consequently, we say the $\boldsymbol{U}$ is a *block-rank-one matrix*.

From the point of view of interpretation, each column of a block-rank-one matrix is the product of a first-level predictor from $\boldsymbol{X}$ and a second-level predictor from $\boldsymbol{Z}$. Because generally both $\boldsymbol{X}$ and $\boldsymbol{Z}$ include an intercept, i.e., a column with all elements equal to 1, this means that the columns of $\boldsymbol{X}$ and $\boldsymbol{Z}$ themselves also occur in $\boldsymbol{U}$, with $\boldsymbol{Z}$ disaggregated. Thus, SOM models have predictors with fixed regression coefficients that are *interactions*, and much of the classical literature on interaction in the linear model, such as Cox [23] and Aiken and West [1], applies to these models as well.

There is one additional consequence of the structure (1.8). We can write

$$[\boldsymbol{U}\boldsymbol{\gamma}]_{ij} = \sum_{s=1}^{p} x_{ijs}\boldsymbol{z}_j'\boldsymbol{\gamma}_s = \sum_{s=1}^{p}\sum_{v=1}^{q} x_{ijs}\gamma_{sv}z_{jv}.$$

Now define the *balanced case* of SOM, in which all $\boldsymbol{X}_j$ are the same. This seems very far fetched if we are thinking of students in classes, but it is actually quite natural for repeated measures. There $\boldsymbol{X}$ could be a basis of growth functions, such as polynomials or exponentials. If measurements are made at the same time points, then indeed all $\boldsymbol{X}_j$ are the same. Other situations in which this may happen are medical or biological experiments, in which dosages of drugs or other treatment variables could be the same, or psychological experiments, in which the stimuli presented to all participants are the same.

In the balanced case, we can rearrange SOM as

$$\underline{\boldsymbol{Y}} = \boldsymbol{Z}\boldsymbol{\Gamma}\boldsymbol{X}' + \underline{\boldsymbol{\Delta}}\boldsymbol{X}' + \underline{\boldsymbol{E}},$$

where the (j,i)-th element of $\underline{\boldsymbol{Y}}$ is $\underline{y}_{ij}$, the j-th row of $\boldsymbol{Z}$ is $\boldsymbol{z}_j'$, the j-th row of $\underline{\boldsymbol{\Delta}}$ is $\underline{\boldsymbol{\delta}}_j'$, the s-th column of $\boldsymbol{\Gamma}$ is $\boldsymbol{\gamma}_s$, and the meaning of the other symbols follows. Thus, the rows are independent. This shows that SOM in this case is a random coefficient version of the classical growth curve model of Potthoff and Roy [91]. Conversely, SOM can be seen as a far-reaching generalization of these classical fixed-effect growth models.

1.4.4 Multilevel Models

Most of the classical multilevel literature, with its origins in education and sociology, deals with the SOM. But in more recent literature, multilevel analysis can refer to more general *Hierarchical Linear Models* or HLMs, of which the two-level MLM (1.4) and the two-level RCM (1.5) are examples. A good example of this more general use, which we also follow throughout the Handbook, is the discussion in Gelman [40].

1.4.5 Generalizations

We shall be very brief about the various generalizations of the multilevel model, because most of these are discussed extensively in the subsequent chapters of this Handbook.

Heteroskedasticity and Conditional Intragroup Dependence

Heteroskedasticity is the phenomenon that residual variances are different for different units. More specifically, it usually means that the variance of the residual depends in some way on the explanatory variables. Heteroskedasticity is a frequently occurring phenomenon in cross-sectional data analysis (and some forms of time series analysis, in particular financial time series). Therefore, we may expect that heteroskedasticity will also be prevalent in many multilevel data analyses. This is indeed the case. In fact, heteroskedasticity is an explicit part of most multilevel models. For example, in the model that we focus on, the covariance matrix of the dependent variables for the j-th group, $\underline{\boldsymbol{y}}_j$, is $\boldsymbol{V}_j = \boldsymbol{X}_j\boldsymbol{\Omega}\boldsymbol{X}_j' + \sigma_j^2\boldsymbol{I}$. Clearly, this depends on $\boldsymbol{X}_j$, so if $\boldsymbol{X}_j$ contains more than just the constant and the corresponding elements of $\boldsymbol{\Omega}$ are not restricted to zero, this induces heteroskedasticity. Furthermore, allowing different residual variances σ_j^2 is also a form of heteroskedasticity.

However, in this specification, the residual variances within the same group are the same, i.e., $\mathrm{Var}(\underline{\epsilon}_{ij}) = \sigma_j^2$, which is the same for all i. Thus, there is

heteroskedasticity between groups, but not within groups. This may be unrealistic in many applications. In such cases, one may want to specify an extended model that explicitly includes within-groups heteroskedasticity. Such a model, and how it can be used to detect heteroskedasticity and thus misspecification of the random part of the model, is described in Chapter 3.

Another widespread phenomenon is lack of independence of observations. Again, this is one of the features of a typical multilevel model: It is assumed that observations within groups are dependent. This gives rise to the well-known intraclass correlation. As we have seen, this is modeled in a typical multilevel model through the random coefficients and, more specifically, through the random terms $\underline{\boldsymbol{\delta}}_j$ in our model specification. However, again this feature does not extend to conditional within-groups comparisons. The units are assumed conditionally independent within their groups, reflected in the diagonality of the covariance matrix of $\underline{\boldsymbol{\epsilon}}_j$. This assumption may also not always be realistic. The leading example in which it is likely to be violated is in longitudinal (or *panel*) data, where the within-groups observations are different observations of the same subject (or object) over time. In such data, residuals often show considerable autocorrelation; i.e., there is a high correlation between residuals that are not far apart. This phenomenon, and how it can be modeled, is discussed extensively in Chapter 7. A similar situation is encountered with *spatial* data, such as data on geographic regions. Then there tends to be spatial autocorrelation; i.e., neighboring regions are "more similar" than regions further apart. See, e.g., Anselin [3] for an overview of modeling spatial autocorrelation. This type of model was integrated in a multilevel model with random coefficients by Elhorst and Zeilstra [37].

More Levels and Different Dependence Structures

Slopes-as-outcomes models can be generalized quite easily to more than two levels. One problem is, however, that matrix notation does not work any more. Switching to scalar notation, we indicate how to generalize by giving the multi-level model for student i_1 in class i_2 in school i_3, and so on. For a model with L levels, it is

$$\underline{\beta}^{(v)}_{i_v,\ldots,i_{v+L-1}} = \sum_{i_{v+L}=1}^{p^{(v)}_{L+1,\ldots,v+L-1}} x^{(v)}_{i_v,\ldots,i_{v+L}} \underline{\beta}^{(v+1)}_{i_{v+1},\ldots,i_{v+L}} + \underline{\epsilon}^{(v)}_{i_v,\ldots,i_{v+L-1}},$$

where superscripts in parentheses indicate the level of the variable. In order to complete the model, we have to assume something about the boundary cases. For level $v = 1$, $\underline{\beta}_{i_1,\ldots,i_L}$ is what we previously wrote as $\underline{y}_{ij}$ for a two-level model, i.e., the value of the outcome for student ij. For the highest level $(L + 1)$, the random coefficients are set to fixed constants, because otherwise

we would have to go on making further specifications. Although the notation becomes somewhat unwieldy, the idea is simple enough.

Other types of different dependence structures are cross-classifications and multiple membership classifications. In the former, an observation is nested in two or more higher-level units, but these higher-level units are not nested within each other. An example is a sample of individuals who are nested within the primary schools and secondary schools that they attended, but not all students from a primary school necessarily attended the same secondary school or vice versa. Multiple membership classifications occur when observations are nested within multiple higher-level units of the same type. For example, patients can be treated by several nurses. These two types of dependency structure are discussed at length in Chapter 8. The notation that is used in that chapter can also be applied to "ordinary" (i.e., nested) multiple-level models, somewhat reducing the unwieldiness mentioned above.

Nonlinear Mixed Models

Nonlinear mixed models come in two flavors. And of course, these nonlinear generalizations specialize in the obvious way to random coefficient and slopes-as-outcomes models.

First, we have *nonlinear mixed models* in which the linear combinations of the predictors are replaced by nonlinear parametric functions, both for the fixed part and the random part. An obvious variation, to reduce the complexity, is to use a nonlinear combination of linear combinations. These nonlinear mixed models are usually fitted with typical nonlinear regression techniques; i.e., we linearize the model around the current estimate and then use linear multilevel techniques. For details we refer to Pinheiro and Bates [89]. Detection and nonparametric modeling of nonlinearities in the fixed part of the model is discussed in more detail in Chapter 3.

Second, we have *generalized linear mixed models*. In the same way as the generalized linear model extends the linear model, the generalized linear mixed model extends the mixed linear model. The basic trick is (in the two-level case) to condition on the random effects and to assume a generalized linear model for the conditional distribution of the outcomes. Then the full model is obtained by multiplying the conditional density by the marginal density of the random effects and integrating. This is, of course, easier said than done, because the high-dimensional integrals that are involved cannot be evaluated in closed form. Thus, sophisticated approximations and algorithms are needed. These are discussed in many of the subsequent chapters, in particular Chapters 2, 5, and 9.

The leading case of applications of nonlinear models is the modeling of nominal and ordinal categorical dependent variables. Several competing spec-

ifications exist, and each has its advantages and disadvantages. These are discussed and compared in detail in Chapter 6.

Multivariate Models, Endogeneity, Measurement Errors, and Latent Variables

In this chapter, we focus on models with one dependent variable, called $\underline{y}$, and explanatory variables (generically called x and z) that are assumed to be fixed constants. Instead of the latter, we can also assume that the explanatory variables are strictly *exogenous* random variables and then do our analysis conditionally on their realizations. This does not change the treatment, the results, or the notation.

In fact, most of the multilevel literature is based on a similar setup, so in that sense this chapter reflects the mainstream of multilevel analysis. In many practical situations, however, this setup is not sufficient, or even clearly incorrect, and extensions or modifications are needed. Here, we briefly mention a few such topics that are somewhat related.

Of these, multiple dependent variables are often most easily accommodated. In most situations, one can simply estimate the models for each of these dependent variables separately. If the different equations do not share any parameters and the dependent variable of one equation does not enter another as explanatory variable, this should be sufficient. Also, as mentioned earlier, multivariate models can be viewed as univariate models with an additional level and thus be estimated within a relatively standard multilevel modeling setup.

Endogeneity is the situation where (at least) one of the explanatory variables in a regression equation is a random variable that is correlated with the error term in the equation of interest. Statistically, this leads to biased and inconsistent estimators. Substantively, this is often the result of one or more unobserved variables that influence both the explanatory variable and the dependent variable in the equation. If it is only considered a statistical nuisance, consistent estimators can usually be obtained by using some form of instrumental variables method [e.g., 126], which has been developed for multilevel analysis by Kim and Frees [65]. In many cases, however, it is of some substantive interest to model the dependence more extensively. Examples of such models are especially abundant in longitudinal situations. Chapter 7 discusses these in detail.

A special source of endogeneity that occurs frequently in the social sciences is measurement error in an explanatory variable. Almost all psychological test scores can be considered as, at best, imperfect measures of some concept that one tries to measure. A notorious example from economics is income. Let us assume that true (log) consumption $\underline{c}^*$ of a household depends on true (log) household income $\underline{y}^*$ through a simple linear regression equation, but the

measurements $\underline{c}$ and $\underline{y}$ of consumption and income are only crude estimates. In formulas,

$$\begin{aligned}\underline{c}^* &= \beta_1 + \beta_2 \underline{y}^* + \underline{\epsilon},\\ \underline{c} &= \underline{c}^* + \underline{v},\\ \underline{y} &= \underline{y}^* + \underline{w},\end{aligned}$$

where we assume that the error terms $\underline{\epsilon}$, $\underline{v}$, and $\underline{w}$ are all mutually independent and independent of $\underline{y}^*$, and we have omitted the indices denoting the observations. We can write the model in terms of the observed variables as

$$\underline{c} = \beta_1 + \beta_2 \underline{y} + \underline{u},$$

where $\underline{u} = \underline{\epsilon} + \underline{v} - \beta_2 \underline{w}$. Because $\underline{w}$ is part of both the explanatory variable $\underline{y}$ and the error term $\underline{u}$, these two are correlated and thus we have the endogeneity problem. An extensive general treatment of measurement error, its statistical consequences, and how to obtain suitable estimators, is given by Wansbeek and Meijer [123]. Goldstein [46, Chapter 13] discusses the handling of measurement errors in multilevel models.

Models that include measurement errors explicitly are a subset of *latent variable models*. Latent variable models typically specify a relationship between substantive concepts, the *structural model*, and a relationship between these concepts and the observed variables (the indicators), which is the *measurement model*. The concepts may be fairly concrete, like income above, but may also be highly abstract theoretical concepts, like personality traits. Most latent variable models are members of the class of *structural equation models*. Because of the flexibility in selecting (multiple) observed variables to analyze and the flexibility in defining latent variables, structural equation models encompass a huge class of models. In particular, multivariate models, endogeneity, measurement errors, and latent variables can all be combined into a single structural equation model. Structural equation models for multilevel data are described extensively in Chapter 12.

Nonnormality

It is customary to specify normal distributions for the random terms in a multilevel model. A normality assumption for error terms can typically be defended by arguing that the error term captures many small unobserved influences, and a central limit theorem then implies that it should be approximately normally distributed. However, normality of random coefficients is often not at all logical. Empirically, in effectiveness studies of schools, hospitals, etc., we might find that many perform "average", whereas there are a few that perform exceptionally well or exceptionally poor. Such a pattern would suggest

a distribution with heavy tails or a mixture distribution. Moreover, the normal distribution has positive density for both positive and negative values, whereas in many cases, theory or common sense (which often coincide) says that a coefficient should have a specific sign. For example, in economics, a higher price should decrease (indirect) utility, and in education, higher intelligence should lead to higher scores on school tests.

In economics, marketing, and transportation, the lognormal distribution has been proposed as a convenient alternative distribution for random coefficients in discrete choice models, perhaps after changing the sign of the explanatory variable. Meijer and Rouwendal [83] discuss this literature and compare normal, lognormal, and Gamma distributions, as well as a nonparametric alternative. In their travel preference data, lognormal and Gamma clearly outperform normal and nonparametric, on the basis of fit and interpretability. Chapter 7 further discusses the nonparametric maximum likelihood estimator.

For *linear* multilevel models, it is fairly straightforward that all the usual estimators are still consistent if the random terms are nonnormally distributed [121]. The standard errors of the fixed coefficients are still correct under nonnormality, but standard errors of the variance parameters must be adjusted. This can be done by using a robust covariance matrix, which will be discussed in Section 1.6.3 below, or by using resampling techniques specifically developed for multilevel data (see Chapter 11).

Estimators of *nonlinear* multilevel analysis models are inconsistent if the distribution of the random coefficients is misspecified. Robust covariance matrices and resampling can give asymptotically correct variability estimators, but it may be questionable whether these are useful if it is unclear whether the estimators of the model parameters are meaningful under gross misspecification of the distributions.

An interesting logical consequence of the line of reasoning that leads to nonnormal distributions is that it also suggests that in cases where the coefficient should have a specific sign, the functional form of the level-2 model should also change. For example, if a level-1 random coefficient $\underline{\beta}$ should be positive, then a specification $\underline{\beta} = \boldsymbol{z}'\boldsymbol{\gamma} + \underline{\delta}$, even with nonnormal $\underline{\delta}$, may be problematic, and a specification

$$\log \underline{\beta} = \boldsymbol{z}'\boldsymbol{\gamma} + \underline{\delta}$$

may make more sense, where now there is nothing wrong with a normal $\underline{\delta}$, because it induces a lognormal $\underline{\beta}$. Remarkably, with this specification, although both level-1 and level-2 submodels are linear in parameters, the combined model is not.

1.5 Loss Functions

Loss functions are used in statistics to measure the badness-of-fit of the model and the given data. In most circumstances, they measure the distance between the observed and the expected values of appropriately chosen statistics such as the means, the dispersions, or the distribution functions. It is quite common in the multilevel literature to concentrate exclusively on the likelihood function or, in a Bayesian context, the posterior density function. We will pay more attention than usual to least squares loss functions, both for historical and didactic reasons.

1.5.1 Least Squares

A general least squares loss function for the multilevel problem (in particular, the SOM) is of the form

$$\rho(\gamma) = \sum_{j=1}^{m} (\boldsymbol{y}_j - \boldsymbol{X}_j \boldsymbol{Z}_j \gamma)' \boldsymbol{A}_j^{-1} (\boldsymbol{y}_j - \boldsymbol{X}_j \boldsymbol{Z}_j \gamma), \tag{1.9}$$

where the weight matrices $\boldsymbol{A}_j$ are supposed to be known (not estimated).

There is a simple trick that can be used to simplify the computations, and to give additional insight into the structure of the loss function. Define the regression coefficients

$$\boldsymbol{b}_j = (\boldsymbol{X}_j' \boldsymbol{A}_j^{-1} \boldsymbol{X}_j)^{-1} \boldsymbol{X}_j' \boldsymbol{A}_j^{-1} \boldsymbol{y}_j$$

and the residuals

$$\boldsymbol{r}_j = \boldsymbol{y}_j - \boldsymbol{X}_j \boldsymbol{b}_j \,.$$

Then $\boldsymbol{y}_j = \boldsymbol{X}_j \boldsymbol{b}_j + \boldsymbol{r}_j$, and $\boldsymbol{X}_j' \boldsymbol{A}_j^{-1} \boldsymbol{r}_j = \emptyset$. Now, for group j,

$$\rho_j(\gamma) = (\boldsymbol{b}_j - \boldsymbol{Z}_j \gamma)' \boldsymbol{X}_j' \boldsymbol{A}_j^{-1} \boldsymbol{X}_j (\boldsymbol{b}_j - \boldsymbol{Z}_j \gamma) + \boldsymbol{r}_j' \boldsymbol{A}_j^{-1} \boldsymbol{r}_j. \tag{1.10}$$

This expression of the loss function is considerably more convenient than (1.9), because it involves smaller vectors and matrices.

If we choose $\boldsymbol{A}_j$ of the form $\boldsymbol{V}_j = \boldsymbol{X}_j \boldsymbol{\Omega} \boldsymbol{X}_j' + \sigma_j^2 \boldsymbol{I}$, again with $\boldsymbol{\Omega}$ and σ_j^2 assumed known, then we can simplify the loss function some more, using the matrix results in Appendix 1.C. Let $\boldsymbol{P}_j \triangleq \boldsymbol{X}_j (\boldsymbol{X}_j' \boldsymbol{X}_j)^{-1} \boldsymbol{X}_j'$, and $\boldsymbol{Q}_j \triangleq \boldsymbol{I} - \boldsymbol{P}_j$. We will also write, in the sequel,

$$\boldsymbol{W}_j \triangleq \boldsymbol{\Omega} + \sigma_j^2 (\boldsymbol{X}_j' \boldsymbol{X}_j)^{-1}.$$

Observe that if $\underline{\boldsymbol{b}}_j \triangleq (\boldsymbol{X}_j' \boldsymbol{X}_j)^{-1} \boldsymbol{X}_j' \underline{\boldsymbol{y}}_j$, then $\boldsymbol{W}_j$ is the dispersion of $\underline{\boldsymbol{b}}_j$. Accordingly, from now on we redefine $\boldsymbol{b}_j \triangleq (\boldsymbol{X}_j' \boldsymbol{X}_j)^{-1} \boldsymbol{X}_j' \boldsymbol{y}_j$ and $\boldsymbol{r}_j \triangleq \boldsymbol{y}_j - \boldsymbol{X}_j \boldsymbol{b}_j$, regardless of the definition of $\boldsymbol{A}_j$.

From Theorem 1.2 in the appendix,

$$\boldsymbol{V}_j^{-1} = \boldsymbol{X}_j(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{W}_j^{-1}(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j' + \sigma_j^{-2}\boldsymbol{Q}_j, \tag{1.11}$$

and thus

$$\boldsymbol{r}_j'\boldsymbol{V}_j^{-1}\boldsymbol{r}_j = \sigma_j^{-2}\boldsymbol{r}_j'\boldsymbol{r}_j = (n_j - p)s_j^2/\sigma_j^2$$

and

$$\boldsymbol{X}_j'\boldsymbol{V}_j^{-1}\boldsymbol{X}_j = \boldsymbol{W}_j^{-1}.$$

Hence,

$$\rho_j(\boldsymbol{\gamma}) = (\boldsymbol{b}_j - \boldsymbol{Z}_j\boldsymbol{\gamma})'\boldsymbol{W}_j^{-1}(\boldsymbol{b}_j - \boldsymbol{Z}_j\boldsymbol{\gamma}) + (n_j - p)s_j^2/\sigma_j^2. \tag{1.12}$$

Computing least squares loss in this way is even more efficient than using (1.10).

1.5.2 Full Information Maximum Likelihood (FIML)

The least squares approach supposes that the weight matrix is known, but, of course, in a more general case the weight function will depend on some unknown parameters that have to be estimated from the same data as the regression coefficients. In that case, we need a loss function that not only measures how close the fitted regression coefficients are to their expected values, but also measures, at the same time, how well the fitted dispersion matrices correspond with the dispersion of the residuals. For this we use the log-likelihood.

As is well known, the method of maximum likelihood has a special position in statistics, especially in applied statistics. Maximum likelihood estimators are introduced as if they are by definition optimal, in all situations. Another peculiarity of the literature is that maximum likelihood methods are introduced by assuming a specific probability model, which is often quite obviously false in the situations one has in mind. In our context, this means that typically it is assumed that the disturbances, and thus the observed $\boldsymbol{y}$, are realizations of jointly normal random variables. Of course, such an assumption is highly debatable in many educational research situations, and quite absurd in others.

Consequently, we take a somewhat different position. Least squares estimates are obtained by minimizing a given loss function. Afterward, we derive their properties and we discover that they behave nicely in some situations. We approach multinormal maximum likelihood in a similar way. The estimates are defined as those values of $\boldsymbol{\gamma}$, $\boldsymbol{\Omega}$, and $\{\sigma_j^2\}$ that minimize the loss function

$$\mathcal{L}^F(\boldsymbol{\gamma}, \boldsymbol{\Omega}, \{\sigma_j^2\}) \triangleq \log|\boldsymbol{V}| + (\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\gamma})'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\gamma}). \tag{1.13}$$

This loss function, which is the negative logarithm of the likelihood function (except for irrelevant constants), is often called the *deviance*. The important

fact here is not that we assume multivariate normality but that (1.13) defines quite a natural loss function. It measures closeness of $\boldsymbol{y}$ to $\boldsymbol{U\gamma}$ by weighted least squares, and it measures at the same time closeness of $\boldsymbol{R}(\boldsymbol{\gamma}) \triangleq (\boldsymbol{y} - \boldsymbol{U\gamma})(\boldsymbol{y} - \boldsymbol{U\gamma})'$ to $\boldsymbol{V}$.

This last property may not be immediately apparent from the form of (1.13). It follows from the inequality $\log|\boldsymbol{A}| + \operatorname{tr}\boldsymbol{A}^{-1}\boldsymbol{B} \geq \log|\boldsymbol{B}| + m$, which is true for all pairs of positive definite matrices of order m. We have equality if and only if $\boldsymbol{A} = \boldsymbol{B}$. Thus, in our context, $\log|\boldsymbol{V}| + \operatorname{tr}\boldsymbol{V}^{-1}\boldsymbol{R}(\boldsymbol{\gamma})$ measures the distance between $\boldsymbol{V}$ and the residuals $\boldsymbol{R}(\boldsymbol{\gamma})$. We want to make residuals small, and we want the dispersion to be maximally similar to the dispersion of the residuals. Moreover, we want to combine these two objectives in a single loss function.

To find simpler expressions for the inverse and the determinant in (1.13), we use the matrix results in Appendix 1.C, in the same way as they were used in Section 1.5.1. From Theorem 1.1 in the appendix,

$$\log|\boldsymbol{V}_j| = (n_j - p)\log\sigma_j^2 + \log|\boldsymbol{X}_j'\boldsymbol{X}_j| + \log|\boldsymbol{W}_j|.$$

If we combine this with result (1.12), we find for group j, ignoring terms that do not depend on the parameters,

$$\mathcal{L}_j^F(\boldsymbol{\gamma}, \boldsymbol{\Omega}, \sigma_j^2) = (n_j - p)\big(\log\sigma_j^2 + s_j^2/\sigma_j^2\big) + \log|\boldsymbol{W}_j| + (\boldsymbol{b}_j - \boldsymbol{Z}_j\boldsymbol{\gamma})'\boldsymbol{W}_j^{-1}(\boldsymbol{b}_j - \boldsymbol{Z}_j\boldsymbol{\gamma}).$$

To distinguish the resulting estimators explicitly from the REML estimators below, these ML estimators are called *full information maximum likelihood* (FIML) in this chapter.

1.5.3 Residual Maximum Likelihood (REML)

In the simplest possible linear model $\underline{y}_i = \mu + \underline{\epsilon}_i$, with $\underline{\epsilon}_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, the maximum likelihood estimator of μ is the mean and that of σ^2 is the sum of squares around the mean, divided by the number of observations n. This estimate of the variance is biased and, as a consequence, the sample variance is usually defined by dividing the sum of squares by $n-1$. The same reasoning, adjusting for bias, in the linear regression model leads to dividing the residual sum of squares by $n - s$, where s is the number of predictors.

We can also arrive at these bias adjustments in a slightly different way, which allows us to continue to use the log-likelihood. Suppose we compute the likelihood of the deviations of the mean, or in the more general case the likelihood of the observed regression residuals. These residuals have a singular multivariate normal distribution, and the maximum likelihood estimate of the variance turns out to be precisely the bias-adjusted estimate. Thus, in

these simple cases, *residual maximum likelihood* (REML; also frequently called *restricted maximum likelihood*) estimates can actually be computed from full information maximum likelihood estimates by a simple multiplicative bias adjustment.

In multilevel models, or more generally in MLMs, bias adjustment is not that easy, but we can continue to use the same reasoning as in the simpler cases and then expect to get an estimator with smaller bias. Let us start with the MLM $\underline{\boldsymbol{y}} = \boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{X}\underline{\boldsymbol{\delta}} + \underline{\boldsymbol{\epsilon}}$. Suppose $\boldsymbol{U}$ is $n \times s$ and of full column rank. Also suppose $\boldsymbol{K}$ is any orthonormal basis for the orthogonal complement of the column space of $\boldsymbol{U}$; i.e., $\boldsymbol{K}$ is an $n \times (n-s)$ matrix with $\boldsymbol{K}'\boldsymbol{K} = \boldsymbol{I}$ and $\boldsymbol{K}'\boldsymbol{U} = \emptyset$. Then define the residuals $\underline{\boldsymbol{r}} \overset{\Delta}{=} \boldsymbol{K}'\underline{\boldsymbol{y}} \sim \mathcal{N}(\emptyset, \boldsymbol{K}'\boldsymbol{V}\boldsymbol{K})$. Thus, the negative loglikelihood or deviance of a realization of $\underline{\boldsymbol{r}}$ is, ignoring the usual constants,

$$\mathcal{L}^R(\boldsymbol{\Omega}, \{\sigma_j^2\}) = \log|\boldsymbol{K}'\boldsymbol{V}\boldsymbol{K}| + \boldsymbol{r}'(\boldsymbol{K}'\boldsymbol{V}\boldsymbol{K})^{-1}\boldsymbol{r}.$$

Observe that this is no longer a function of $\boldsymbol{\gamma}$. Thus, we cannot compute maximum likelihood estimates of the fixed regression coefficients by minimizing this loss function.

Now use Theorem 1.3 from Appendix 1.C, which shows that

$$\boldsymbol{r}'(\boldsymbol{K}'\boldsymbol{V}\boldsymbol{K})^{-1}\boldsymbol{r} = \min_{\boldsymbol{\gamma}}(\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\gamma})'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\gamma}).$$

Harville [52] shows that

$$\log|\boldsymbol{K}'\boldsymbol{V}\boldsymbol{K}| = \log|\boldsymbol{V}| + \log|\boldsymbol{U}'\boldsymbol{V}^{-1}\boldsymbol{U}| - \log|\boldsymbol{U}'\boldsymbol{U}|$$

and, consequently, except for irrelevant constants,

$$\mathcal{L}^R(\boldsymbol{\Omega}, \{\sigma_j^2\}) = \log|\boldsymbol{U}'\boldsymbol{V}^{-1}\boldsymbol{U}| + \min_{\boldsymbol{\gamma}} \mathcal{L}^F(\boldsymbol{\gamma}, \boldsymbol{\Omega}, \{\sigma_j^2\}).$$

It follows that the loss functions for FIML and REML only differ by the term $\log|\boldsymbol{U}'\boldsymbol{V}^{-1}\boldsymbol{U}|$, which can be thought of as a bias correction. In SOM, we can use

$$\boldsymbol{U}'\boldsymbol{V}^{-1}\boldsymbol{U} = \sum_{j=1}^m \boldsymbol{Z}_j'\boldsymbol{W}_j^{-1}\boldsymbol{Z}_j,$$

and, if (1.8) applies, then

$$\boldsymbol{U}'\boldsymbol{V}^{-1}\boldsymbol{U} = \sum_{j=1}^m \boldsymbol{W}_j^{-1} \otimes \boldsymbol{z}_j\boldsymbol{z}_j'.$$

1.5.4 Bayesian Multilevel Analysis

In the Bayesian approach to multilevel analysis, the parameters are treated as random variables, so in our notation they would be written as $\underline{\boldsymbol{\gamma}}$, $\underline{\boldsymbol{\Omega}}$, and

$\{\underline{\sigma}_j^2\}$, jointly denoted as $\underline{\boldsymbol{\theta}}$. Then a *prior distribution* for $\underline{\boldsymbol{\theta}}$ is specified, which is completely known. The parameters of this prior distribution are called *hyperparameters* and their values reflect the state of knowledge about $\underline{\boldsymbol{\theta}}$. In the absence of prior knowledge, this typically means that variances of the parameters are chosen to be infinite or at least very large. Given the specification of the prior distribution, the *posterior distribution* of $\underline{\boldsymbol{\theta}}$, given the observed sample, is found by application of Bayes' theorem:

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{f(\boldsymbol{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{y})} = Cf(\boldsymbol{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}),$$

where $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ is the posterior density, $\pi(\boldsymbol{\theta})$ is the specified prior density, $f(\boldsymbol{y} \mid \boldsymbol{\theta})$ is the conditional normal density that we have been using all along (which is equal to the likelihood function), and C is a normalizing constant that does not depend on $\boldsymbol{\theta}$. An explicit expression for C is rarely needed. The posterior density contains all information about $\underline{\boldsymbol{\theta}}$; all inferences about $\underline{\boldsymbol{\theta}}$ are derived from it. It combines the prior information and the information contained in the sample in a sound (and optimal) way.

From this description, it appears that the Bayesian approach does not fit into our framework of specifying a loss function and then optimizing it. However, in the Bayesian approach, it is common to use the posterior mode or posterior mean as an "estimator" and to compute intervals that contain $100(1-\alpha)\%$ (e.g., 95%) of the probability mass, which act as a kind of "confidence interval". The posterior mean $\hat{\mu}_g$ of parameter g is the argument for which the loss function $E[(\underline{\theta}_g - \mu_g)^2]$, where the expectation is taken over the posterior distribution, attains its minimum, whereas the posterior mode $\hat{\boldsymbol{\theta}}_M$ is by definition the value for which the posterior density $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ attains its maximum or, equivalently, the loss function $-p(\boldsymbol{\theta} \mid \boldsymbol{y})$ attains its minimum. Both are very natural loss functions and, thus, in this way the Bayesian approach neatly fits within our framework. An important advantage of the Bayesian "confidence intervals", especially for the variance parameters, is that they may be asymmetric, reflecting a nonnormal posterior distribution. This is often more realistic for the variance parameters in small to moderate samples.

An important reason for the increasing popularity of the Bayesian approach is that it is able to deal with nonlinear models in a fairly straightforward way, using *Markov chain Monte Carlo* (MCMC) techniques. This gives good results where non-Bayesian approaches often have great difficulty in obtaining good estimators. Chapter 2 is an extensive discussion of the Bayesian approach, and in several other chapters, especially those dealing with nonlinear models, it is also discussed, applied, and compared to likelihood-based approaches. Therefore, we will not discuss it in more detail in this chapter.

1.5.5 Missing Data

It is implicit in the discussion thus far that we have assumed that there are no missing data. In practice, the fact that there are missing data is a widespread phenomenon and often a problem. We can distinguish between *unit nonresponse*, in which no information is available for a targeted observation, and *item nonresponse*, where information is available for some variables but not for others. If we assume that unit nonresponse is not related to any of the random variables ($\underline{\boldsymbol{\delta}}$, $\underline{\boldsymbol{\epsilon}}$) of interest for the missing unit, we can simply proceed by analyzing the observed data set. If it is suspected that unit nonresponse leads to distortions, weighting can be applied (and is often applied) to let the sample distribution of some key variables match the (assumed known) population distribution. See Section 1.8 below for a discussion of sampling weights in multilevel models.

With *item nonresponse*, the simplest and most frequently applied solution is to simply omit all observations for which one or more variables are missing (*listwise deletion*). Although widely used, it is generally considered a bad method. It omits useful information and thus gives inefficient estimators. Even more importantly, it may easily lead to biases in the analyses, if the missing data patterns are related to the variables of interest. Chapter 10 extensively discusses how missing data can be treated in a sound and systematic way.

1.6 Techniques and Algorithms

If we have a loss function, then the obvious associated technique to estimate parameters is to minimize the loss function. Of course, for nonlinear optimization problems there are many different minimization methods. Some are general-purpose optimization methods that can be applied to any multivariate function, and some take the properties of the loss function explicitly into account.

1.6.1 Ordinary and Weighted Least Squares

As we have see in a previous section, the SOM model can be expressed in two steps, as in

$$\underline{\boldsymbol{y}}_j = \boldsymbol{X}_j\underline{\boldsymbol{\beta}}_j + \underline{\boldsymbol{\epsilon}}_j, \tag{1.14a}$$

$$\underline{\boldsymbol{\beta}}_j = \boldsymbol{Z}_j\boldsymbol{\gamma} + \underline{\boldsymbol{\delta}}_j, \tag{1.14b}$$

or in a single-step, as in

$$\underline{\boldsymbol{y}}_j = \boldsymbol{X}_j\boldsymbol{Z}_j\boldsymbol{\gamma} + \boldsymbol{X}_j\underline{\boldsymbol{\delta}}_j + \underline{\boldsymbol{\epsilon}}_j. \tag{1.15}$$

The one-step (1.15) and the two-step (1.14) specifications of the multilevel model suggest two different ordinary least squares methods for fitting the model. This was already discussed in detail by Boyd and Iversen [11]. We follow the treatment of de Leeuw and Kreft [28].

The two-step method first estimates the $\boldsymbol{\beta}_j$ by

$$\boldsymbol{b}_j = (\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j'\boldsymbol{y}_j, \tag{1.16}$$

and then $\boldsymbol{\gamma}$ by

$$\hat{\boldsymbol{\gamma}} = \left(\sum_{j=1}^{m} \boldsymbol{Z}_j'\boldsymbol{Z}_j\right)^{-1} \sum_{j=1}^{m} \boldsymbol{Z}_j'\boldsymbol{b}_j. \tag{1.17}$$

Within the framework of Section 1.5.1, this is obtained by choosing $\boldsymbol{A}_j = \boldsymbol{X}_j\boldsymbol{X}_j' + \boldsymbol{Q}_j$, so that $\boldsymbol{A}_j^{-1} = \boldsymbol{X}_j(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-2}\boldsymbol{X}_j' + \boldsymbol{Q}_j$.

The one-step method estimates $\boldsymbol{\gamma}$ directly from (1.15) as

$$\hat{\boldsymbol{\gamma}} = \left(\sum_{j=1}^{m} \boldsymbol{Z}_j'\boldsymbol{X}_j'\boldsymbol{X}_j\boldsymbol{Z}_j\right)^{-1} \sum_{j=1}^{m} \boldsymbol{Z}_j'\boldsymbol{X}_j'\boldsymbol{y}_j.$$

By using (1.16), we see immediately, however, that the one-step method can also be written as

$$\hat{\boldsymbol{\gamma}} = \left(\sum_{j=1}^{m} \boldsymbol{Z}_j'\boldsymbol{X}_j'\boldsymbol{X}_j\boldsymbol{Z}_j\right)^{-1} \sum_{j=1}^{m} \boldsymbol{Z}_j'\boldsymbol{X}_j'\boldsymbol{X}_j\boldsymbol{b}_j. \tag{1.18}$$

Thus, the one-step estimate can be computed in two steps as well. Within the framework of Section 1.5.1, the one-step estimate is obtained by choosing $\boldsymbol{A}_j = \boldsymbol{I}$.

Both methods provide unbiased estimators of $\boldsymbol{\gamma}$, they are non-iterative, and they are easy to implement. An expression for their dispersion matrices is easily obtained by using $\mathrm{Cov}(\underline{\boldsymbol{b}}_j) = \boldsymbol{W}_j$, which was obtained above. Hence, the dispersion matrix of the two-step estimator is

$$\left(\sum_{j=1}^{m} \boldsymbol{Z}_j'\boldsymbol{Z}_j\right)^{-1} \left(\sum_{j=1}^{m} \boldsymbol{Z}_j'\boldsymbol{W}_j\boldsymbol{Z}_j\right) \left(\sum_{j=1}^{m} \boldsymbol{Z}_j'\boldsymbol{Z}_j\right)^{-1}$$

and the dispersion matrix of the one-step estimator is

$$\left(\sum_{j=1}^{m} \boldsymbol{Z}_j'\boldsymbol{X}_j'\boldsymbol{X}_j\boldsymbol{Z}_j\right)^{-1} \left(\sum_{j=1}^{m} \boldsymbol{Z}_j'\boldsymbol{X}_j'\boldsymbol{X}_j\boldsymbol{W}_j\boldsymbol{X}_j'\boldsymbol{X}_j\boldsymbol{Z}_j\right) \left(\sum_{j=1}^{m} \boldsymbol{Z}_j'\boldsymbol{X}_j'\boldsymbol{X}_j\boldsymbol{Z}_j\right)^{-1}.$$

Despite their virtues, these least squares estimators have fallen into disgrace in the mainstream multilevel world, because they are neither BLUE nor BLUP

[43, 103]. This is somewhat supported by the simulations reported (for a three-level model) in Cheong et al. [21], where especially for level-1 covariates efficiencies of ML estimators are substantially higher (up to 55%). The one-step OLS estimator still enjoys a great popularity in economics, though.

The next candidate that comes to mind applies if both $\boldsymbol{\Omega}$ and $\{\sigma_j^2\}$ are known. We can then compute the WLS estimate

$$\hat{\boldsymbol{\gamma}} = \left(\sum_{j=1}^m \boldsymbol{Z}_j'\boldsymbol{X}_j'\boldsymbol{V}_j^{-1}\boldsymbol{X}_j\boldsymbol{Z}_j\right)^{-1}\sum_{j=1}^m \boldsymbol{Z}_j'\boldsymbol{X}_j'\boldsymbol{V}_j^{-1}\boldsymbol{y}_j. \tag{1.19}$$

As we have seen, this can be simplified to

$$\hat{\boldsymbol{\gamma}} = \left(\sum_{j=1}^m \boldsymbol{Z}_j'\boldsymbol{W}_j^{-1}\boldsymbol{Z}_j\right)^{-1}\sum_{j=1}^m \boldsymbol{Z}_j'\boldsymbol{W}_j^{-1}\boldsymbol{b}_j. \tag{1.20}$$

Within the framework of Section 1.5.1, the WLS estimate is obtained by choosing $\boldsymbol{A}_j = \boldsymbol{V}_j$. The dispersion matrix of the WLS estimator is obtained analogously to the ones above, and in this case it simplifies to

$$\left(\sum_{j=1}^m \boldsymbol{Z}_j'\boldsymbol{W}_j^{-1}\boldsymbol{Z}_j\right)^{-1}.$$

The formal similarity of (1.17), (1.18), and (1.20) is clear. They can all be thought of as two-step methods, which first compute the $\boldsymbol{b}_j$ and then do a weighted regression of the $\boldsymbol{b}_j$ on the $\boldsymbol{Z}_j$. Of course, (1.20) is mostly useless by itself, because we do not know what σ_j^2 and $\boldsymbol{\Omega}$ are, but we can insert consistent estimators of these instead. A method to compute consistent estimators of the elements of the variance parameters from the OLS residuals is discussed in de Leeuw and Kreft [28], and is also discussed below. The resulting method for estimating $\boldsymbol{\gamma}$ is fully efficient and non-iterative.

For WLS estimators with estimators of the variance parameters inserted, the exact covariance matrix generally cannot be computed. However, it follows from standard large sample theory (Slutsky's theorem; see, e.g., Ferguson [38] or Wansbeek and Meijer [123, pp. 369–370]) that if the estimators of $\boldsymbol{\Omega}$ and σ_j^2 are consistent, then the asymptotic distribution of the WLS estimator of $\boldsymbol{\gamma}$ is the same as the (asymptotic) distribution of the hypothetical estimator (1.20) that uses the true values of $\boldsymbol{\Omega}$ and σ_j^2 in the weight matrix, so we can still use the covariance matrix given above, especially with larger sample sizes.

The BLUE and the BLUP

Consider the model $\underline{\boldsymbol{y}} \sim \mathcal{N}(\boldsymbol{U}\boldsymbol{\gamma}, \boldsymbol{V})$. A linear estimator of the form $\hat{\underline{\boldsymbol{\gamma}}} = \boldsymbol{L}'\underline{\boldsymbol{y}}$ is unbiased if $\boldsymbol{L}'\boldsymbol{U} = \boldsymbol{I}$, and it has dispersion $\boldsymbol{L}'\boldsymbol{V}\boldsymbol{L}$. The dispersion matrix

is minimized, in the Löwner [77] ordering of matrices (i.e., $\boldsymbol{A} \geq \boldsymbol{B}$ if $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite), by choosing $\boldsymbol{L} = \boldsymbol{V}^{-1}\boldsymbol{U}(\boldsymbol{U}'\boldsymbol{V}^{-1}\boldsymbol{U})^{-1}$. Thus,

$$\hat{\underline{\gamma}} = (\boldsymbol{U}'\boldsymbol{V}^{-1}\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{V}^{-1}\underline{\boldsymbol{y}}$$

is the *best linear unbiased estimator* or BLUE. In the SOM,

$$\boldsymbol{U}'\boldsymbol{V}^{-1}\boldsymbol{U} = \sum_{j=1}^{m} \boldsymbol{Z}_j'\boldsymbol{W}_j^{-1}\boldsymbol{Z}_j$$

and

$$\boldsymbol{U}'\boldsymbol{V}^{-1}\underline{\boldsymbol{y}} = \sum_{j=1}^{m} \boldsymbol{Z}_j'\boldsymbol{W}_j^{-1}\underline{\boldsymbol{b}}_j.$$

Thus, the BLUE is given by (1.20).

We can also look at estimates of the error components. Of course, this means we are estimating random variables and, consequently, the *best linear unbiased predictor* or BLUP is a more appropriate term than the BLUE. To find the BLUP, we minimize the mean squared prediction error

$$\text{MSPE} \stackrel{\Delta}{=} E\big[(\boldsymbol{L}'\underline{\boldsymbol{y}} + \boldsymbol{a} - \underline{\boldsymbol{\delta}})(\boldsymbol{L}'\underline{\boldsymbol{y}} + \boldsymbol{a} - \underline{\boldsymbol{\delta}})'\big] \tag{1.21a}$$

over $\boldsymbol{L}$ and $\boldsymbol{a}$ on the condition that

$$E(\boldsymbol{L}'\underline{\boldsymbol{y}} + \boldsymbol{a} - \underline{\boldsymbol{\delta}}) = \boldsymbol{\emptyset}. \tag{1.21b}$$

From (1.21b) we obtain $\boldsymbol{a} = -\boldsymbol{L}'\boldsymbol{U}\boldsymbol{\gamma}$, which means that the mean squared prediction error (1.21a) is

$$\begin{aligned}\text{MSPE} &= \boldsymbol{L}'\boldsymbol{V}\boldsymbol{L} - \boldsymbol{L}'\boldsymbol{X}\boldsymbol{\Omega} - \boldsymbol{\Omega}\boldsymbol{X}'\boldsymbol{L} + \boldsymbol{\Omega} \\ &= (\boldsymbol{V}\boldsymbol{L} - \boldsymbol{X}\boldsymbol{\Omega})'\boldsymbol{V}^{-1}(\boldsymbol{V}\boldsymbol{L} - \boldsymbol{X}\boldsymbol{\Omega}) + \boldsymbol{\Omega} - \boldsymbol{\Omega}\boldsymbol{Z}'\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{\Omega} \\ &\geq \boldsymbol{\Omega} - \boldsymbol{\Omega}\boldsymbol{Z}'\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{\Omega},\end{aligned}$$

with equality if $\boldsymbol{L} = \boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{\Omega}$, i.e., if

$$\hat{\boldsymbol{\delta}} = \boldsymbol{\Omega}\boldsymbol{X}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\gamma}).$$

In the SOM, using (1.11),

$$\hat{\boldsymbol{\delta}}_j = \boldsymbol{\Omega}\boldsymbol{W}_j^{-1}(\boldsymbol{b}_j - \boldsymbol{Z}_j\boldsymbol{\gamma}),$$

and thus

$$\hat{\boldsymbol{\beta}}_j = \boldsymbol{Z}_j\boldsymbol{\gamma} + \hat{\boldsymbol{\delta}}_j = \boldsymbol{\Omega}\boldsymbol{W}_j^{-1}\boldsymbol{b}_j + (\boldsymbol{I} - \boldsymbol{\Omega}\boldsymbol{W}_j^{-1})\boldsymbol{Z}_j\boldsymbol{\gamma}. \tag{1.22}$$

Thus, the BLUP of the random effects is a matrix weighted average [19] of the least squares estimates $\boldsymbol{b}_j$ and the expected values $\boldsymbol{Z}_j\boldsymbol{\gamma}$. The within-group

least squares estimates are shrunken toward the overall model-based estimate $\boldsymbol{Z}_j\boldsymbol{\gamma}$ of the regression coefficients. This shrinking, which is common in BLUP and related empirical Bayes procedures, is also the basis for the discussion of *borrowing strength*, which has played a major role in the multilevel literature [cf. 13, 101].

Of course, (1.22) contains unknown parameters, and in order to use it in practice, we substitute whatever estimates we have for these unknown parameters.

Estimating the Variance Parameters

As we have seen, for the WLS estimator of $\boldsymbol{\gamma}$ and the BLUP of the random effects, we need consistent estimators of σ_j^2 and $\boldsymbol{\Omega}$. Moreover, estimating these parameters is often one of the main goals of a multilevel analysis and the focus on the random effects is perhaps the most salient difference between multilevel analysis and ordinary regression analysis.

A simple unbiased estimator of σ_j^2 is, of course, the within-groups residual variance $\underline{s}_j^2$. Given the assumptions above,

$$(n_j - p)\underline{s}_j^2/\sigma_j^2 \sim \chi^2_{n_j-p},$$

so that in addition to $E(\underline{s}_j^2) = \sigma_j^2$, we also have $\text{Var}(\underline{s}_j^2) = 2(\sigma_j^2)^2/(n_j - p)$. Furthermore, $\underline{s}_j^2$ is independent of $\underline{\boldsymbol{b}}_j$. However, the variance, chi-square distribution, and independence result depend critically on the normality assumption. If all σ_j^2 are assumed equal, then its natural unbiased estimator is

$$\underline{s}^2 \triangleq \frac{1}{n-p}\sum_{j=1}^{m}(n_j - p)\underline{s}_j^2,$$

where n is total sample size. Under the model assumptions,

$$(n-p)\underline{s}^2/\sigma^2 \sim \chi^2_{n-p},$$

so that $E(\underline{s}^2) = \sigma^2$ and $\text{Var}(\underline{s}^2) = 2(\sigma^2)^2/(n-p)$. Note that consistency of $\underline{s}_j^2$ requires $n_j \to \infty$. This is a little problematic because in some standard asymptotic theory for multilevel analysis (e.g., Longford [76, p. 252]; Verbeke and Lesaffre [120, Lemma 3]), it is assumed that the group sizes are bounded. However, close scrutiny of their theories reveals that the general asymptotic theory should still be valid under a hypothetical sequence such that $m \to \infty$, $n_j \to \infty$, and $n_j/m \to 0$. Maybe even weaker assumptions suffice. Of course, with (many) small groups, $n_j \to \infty$ may not be a useful assumption anyway. On the other hand, consistency of $\underline{s}^2$ only requires $n \to \infty$, which is obviously much weaker. However, the latter also requires the much stronger assumption that all residual variances are equal.

Observing that $\boldsymbol{\Omega} = \text{Cov}(\underline{\boldsymbol{\beta}}_j) = E\big[(\underline{\boldsymbol{\beta}}_j - \boldsymbol{Z}_j\boldsymbol{\gamma})(\underline{\boldsymbol{\beta}}_j - \boldsymbol{Z}_j\boldsymbol{\gamma})'\big]$, a simple estimator of $\boldsymbol{\Omega}$ is obtained by inserting the least squares estimators of $\underline{\boldsymbol{\beta}}_j$ and $\boldsymbol{\gamma}$ in this expression:

$$\hat{\underline{\boldsymbol{\Omega}}} = \frac{1}{m}\sum_{j=1}^{m}(\underline{\boldsymbol{b}}_j - \boldsymbol{Z}_j\hat{\underline{\boldsymbol{\gamma}}})(\underline{\boldsymbol{b}}_j - \boldsymbol{Z}_j\hat{\underline{\boldsymbol{\gamma}}})',$$

or perhaps with $m-1$ instead of m in the denominator, and where $\hat{\underline{\boldsymbol{\gamma}}}$ is the one-step or two-step OLS estimator. Such an estimator is used in the MLA program [16] as "least squares estimator" of $\boldsymbol{\Omega}$ and as starting value for the iterations for obtaining the ML estimators. However, this estimator is biased for two reasons: The variability of $\hat{\underline{\boldsymbol{\gamma}}}$ is not taken into account and the covariance matrix of $\underline{\boldsymbol{b}}_j$ is not $\boldsymbol{\Omega}$, but $\boldsymbol{W}_j$. The first cause of bias vanishes as $m \to \infty$ and the second vanishes as $n_j \to \infty$, so it is only a reasonably good estimator if sample sizes at both levels are large. We can compute its exact expectation and exact variances of its elements, but we will not do that here. In addition to its simplicity, however, it has the virtue that it is guaranteed to be positive (semi)definite. This may prevent numerical problems when used as a starting value in an iterative procedure. Kovačević and Rai [66] propose a similar estimator, with $\boldsymbol{Z}_j\hat{\boldsymbol{\gamma}}$ replaced by the sample average of the $\boldsymbol{b}_j$'s, as a "conservative approximation".

Based on earlier formulas of Swamy [117], de Leeuw and Kreft [28] derive an unbiased estimator of $\boldsymbol{\Omega}$. The estimator of $\boldsymbol{\Omega}$ is derived elementwise. Thus, we look at its (k,l)-th element ω_{kl} and define an unbiased estimator of this element. By doing this for all distinct elements of $\boldsymbol{\Omega}$, we obtain an unbiased estimator of $\boldsymbol{\Omega}$.

Consider the k-th element of $\underline{\boldsymbol{\beta}}_j$, $\underline{\beta}_{jk}$. According to the model assumptions,

$$\underline{\beta}_{jk} = \boldsymbol{z}'_{jk}\boldsymbol{\gamma}_k + \underline{\delta}_{jk},$$

where $\boldsymbol{\gamma}_k$ is a subvector of $\boldsymbol{\gamma}$. The corresponding subvector of the two-step OLS estimator $\hat{\underline{\boldsymbol{\gamma}}}$ is $\hat{\underline{\boldsymbol{\gamma}}}_k$. Let $\boldsymbol{Z}_k$ be the $m \times q_k$ matrix with j-th row $\boldsymbol{z}'_{jk}$, where q_k is the number of elements of $\boldsymbol{z}_{jk}$, i.e., the number of explanatory variables for the k-th random coefficient. Correspondingly, let $\underline{\boldsymbol{b}}_k$ be the vector of length m with $\underline{b}_{jk}$ as its j-th element. Then it follows straightforwardly from the derivation of $\hat{\underline{\boldsymbol{\gamma}}}$ and the structure of $\boldsymbol{Z}_j$ that

$$\hat{\underline{\boldsymbol{\gamma}}}_k = (\boldsymbol{Z}'_k\boldsymbol{Z}_k)^{-1}\boldsymbol{Z}'_k\underline{\boldsymbol{b}}_k.$$

Let $\hat{\underline{\boldsymbol{t}}}_k$ be the vector of length m with $\hat{\underline{t}}_{jk} = \underline{b}_{jk} - \boldsymbol{z}'_{jk}\hat{\underline{\boldsymbol{\gamma}}}_k$ as its j-th element. Then we have

$$\hat{\underline{\boldsymbol{t}}}_k = \boldsymbol{Q}_k\underline{\boldsymbol{b}}_k = \boldsymbol{Q}_k(\underline{\boldsymbol{b}}_k - \boldsymbol{Z}_k\boldsymbol{\gamma}_k) = \boldsymbol{Q}_k\underline{\boldsymbol{t}}_k,$$

where $\boldsymbol{Q}_k = \boldsymbol{I}_m - \boldsymbol{Z}_k(\boldsymbol{Z}'_k\boldsymbol{Z}_k)^{-1}\boldsymbol{Z}'_k$ and $\underline{\boldsymbol{t}}_k$ is implicitly defined. Note that $E(\underline{\boldsymbol{b}}_k) = \boldsymbol{Z}_k\boldsymbol{\gamma}_k$ and

$$\text{Cov}(\underline{\boldsymbol{b}}_k,\underline{\boldsymbol{b}}_l')=\bigoplus_{j=1}^m(\boldsymbol{W}_j)_{kl}=\text{diag}[(\boldsymbol{W}_j)_{kl}]=\omega_{kl}\boldsymbol{I}_m+\boldsymbol{\Sigma}\boldsymbol{\nabla}_{kl},$$

where $\boldsymbol{\Sigma}$ is the diagonal matrix with j-th diagonal element equal to σ_j^2 and $\boldsymbol{\nabla}_{kl}$ is the diagonal matrix with j-th diagonal element equal to $[(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}]_{kl}$. It follows that $E(\hat{\underline{\boldsymbol{t}}}_k)=\emptyset$ and

$$E(\hat{\underline{\boldsymbol{t}}}_k\hat{\underline{\boldsymbol{t}}}_l')=\text{Cov}(\hat{\underline{\boldsymbol{t}}}_k,\hat{\underline{\boldsymbol{t}}}_l')=\omega_{kl}\boldsymbol{Q}_k\boldsymbol{Q}_l+\boldsymbol{Q}_k\boldsymbol{\Sigma}\boldsymbol{\nabla}_{kl}\boldsymbol{Q}_l.$$

It is now natural to define the estimator

$$\hat{\underline{\omega}}_{kl}\triangleq\frac{\text{tr}\left[\hat{\underline{\boldsymbol{t}}}_k\hat{\underline{\boldsymbol{t}}}_l'-\boldsymbol{Q}_k\hat{\underline{\boldsymbol{\Sigma}}}\boldsymbol{\nabla}_{kl}\boldsymbol{Q}_l\right]}{\text{tr}(\boldsymbol{Q}_k\boldsymbol{Q}_l)}=\frac{1}{m^*}\left[\hat{\underline{\boldsymbol{t}}}_l'\hat{\underline{\boldsymbol{t}}}_k-\text{tr}(\hat{\underline{\boldsymbol{\Sigma}}}\boldsymbol{\nabla}_{kl}\boldsymbol{Q}_l\boldsymbol{Q}_k)\right],$$

where $m^*=\text{tr}(\boldsymbol{Q}_k\boldsymbol{Q}_l)$ and $\hat{\underline{\boldsymbol{\Sigma}}}$ is the diagonal matrix with j-th diagonal element equal to $\underline{s}_j^2$. This estimator of ω_{kl} is optimal in the least squares sense and it is evidently unbiased. However, unbiasedness in this context is not necessarily good, because it can easily lead to negative variance estimates.

Noticing that $\hat{\underline{\omega}}_{kl}$ is a quadratic function of the data, its variance can be found by using standard results about the expectations of quadratic forms in normally distributed random variables. The resulting expression is

$$\begin{aligned}\text{Var}(\hat{\underline{\omega}}_{kl})=\frac{1}{(m^*)^2}\Bigg\{&\sum_{i=1}^m\sum_{j=1}^m\big[(\boldsymbol{W}_i)_{ll}(\boldsymbol{W}_j)_{kk}(\boldsymbol{Q}_l\boldsymbol{Q}_k)_{ij}^2\\&\qquad+(\boldsymbol{W}_i)_{kl}(\boldsymbol{W}_j)_{kl}(\boldsymbol{Q}_l\boldsymbol{Q}_k)_{ij}(\boldsymbol{Q}_l\boldsymbol{Q}_k)_{ji}\big]\\&+\sum_{j=1}^m\frac{2(\sigma_j^2)^2}{n_j-p}[(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}]_{kl}^2(\boldsymbol{Q}_l\boldsymbol{Q}_k)_{jj}^2\Bigg\}.\end{aligned}$$

An estimator of this variance is obtained by inserting the estimators $\underline{s}_j^2$ for σ_j^2 and $\hat{\underline{\boldsymbol{\Omega}}}$ for $\boldsymbol{\Omega}$ (the latter in $\boldsymbol{W}_j$) in this formula.

A somewhat related but slightly different method for estimating the variance parameters uses the same ideas as the WLS estimator above, but reverses the roles of the fixed coefficients and the variance parameters. In particular, assume that $\boldsymbol{\gamma}$ is known and that $\boldsymbol{\Omega}$ is written in the linear form (1.3). Then

$$\begin{aligned}E\big[(\underline{\boldsymbol{y}}-\boldsymbol{U}\boldsymbol{\gamma})(\underline{\boldsymbol{y}}-\boldsymbol{U}\boldsymbol{\gamma})'\big]&=\boldsymbol{V}\\&=\bigoplus_{j=1}^m(\boldsymbol{X}_j\boldsymbol{\Omega}\boldsymbol{X}_j'+\sigma_j^2\boldsymbol{I}_{n_j})\\&=\sum_{g=1}^G\left(\bigoplus_{j=1}^m\boldsymbol{X}_j\boldsymbol{C}_g\boldsymbol{X}_j'\right)\xi_g+\sum_{j=1}^m(\boldsymbol{e}_j\boldsymbol{e}_j'\otimes\boldsymbol{I}_{n_j})\sigma_j^2,\end{aligned}$$

where $\boldsymbol{e}_j$ is the j-th column of $\boldsymbol{I}_m$, and if all residual variances are equal, the last summation reduces to $\sigma^2\boldsymbol{I}_n$. Clearly, this expectation is linear in the parameters $\{\xi_g\}$ and $\{\sigma_j^2\}$.

Now, let $\boldsymbol{U}^*[n, G+m]$ be the matrix with g-th column equal to

$$\boldsymbol{U}_g^* \stackrel{\Delta}{=} \operatorname{vec}\left(\bigoplus_{j=1}^{m} \boldsymbol{X}_j \boldsymbol{C}_g \boldsymbol{X}_j'\right)$$

and $(G+j)$-th column equal to

$$\boldsymbol{U}_{G+j}^* \stackrel{\Delta}{=} \operatorname{vec}(\boldsymbol{e}_j\boldsymbol{e}_j' \otimes \boldsymbol{I}_{n_j}).$$

Furthermore, let $\boldsymbol{\gamma}^*[G+m]$ be the vector with g-th element ξ_g $(g = 1, \ldots, G)$ and $(G+j)$-th element σ_j^2 $(j = 1, \ldots, m)$. If all σ_j^2 are equal, $\boldsymbol{U}^*$ has $G+1$ columns, the last one being $\operatorname{vec}\boldsymbol{I}_n$, and $\boldsymbol{\gamma}^*$ has $G+1$ elements, the last one being σ^2. The rest of the discussion is unaltered. Finally, let

$$\boldsymbol{y}^* \stackrel{\Delta}{=} \operatorname{vec}\big[(\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\gamma})(\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\gamma})'\big]. \tag{1.23}$$

Then $E\,\underline{\boldsymbol{y}}^* = \boldsymbol{U}^*\boldsymbol{\gamma}^*$, which suggests that the variance parameters in $\boldsymbol{\gamma}^*$ can be jointly estimated by a least squares method. Although an OLS method would be computationally much easier, a WLS method is typically used, for reasons that become clear in Section 1.6.2 below. From the characteristics of the normal distribution, it follows that the dispersion matrix of $\underline{\boldsymbol{y}}^*$ is $2\boldsymbol{N}_n(\boldsymbol{V} \otimes \boldsymbol{V})$ (e.g., Magnus and Neudecker [79, Lemma 9]), where $\boldsymbol{N}_n[n^2, n^2]$ is a symmetric idempotent matrix of rank $n(n+1)/2$, which projects a column vector of order n^2 onto the space of vec's of symmetric matrices. It is therefore called the *symmetrization matrix* by Wansbeek and Meijer [123, p. 361]. Thus, the dispersion matrix of $\underline{\boldsymbol{y}}^*$ is singular, the reason being that $\underline{\boldsymbol{y}}^*$ contains duplicated elements. We can remove the duplicated elements and then compute the nonsingular dispersion matrix and use it in a WLS procedure. Due to the structure of the problem, this is equivalent to computing the estimate

$$\hat{\boldsymbol{\gamma}}^* = \big((\boldsymbol{U}^*)'(\boldsymbol{V}^*)^{-1}(\boldsymbol{U}^*)\big)^{-1}(\boldsymbol{U}^*)'(\boldsymbol{V}^*)^{-1}(\boldsymbol{y}^*), \tag{1.24}$$

where $\boldsymbol{V}^* = 2(\boldsymbol{V} \otimes \boldsymbol{V})$. From the derivation, it follows immediately that

$$\operatorname{Cov}(\underline{\hat{\boldsymbol{\gamma}}}^*) = \big((\boldsymbol{U}^*)'(\boldsymbol{V}^*)^{-1}(\boldsymbol{U}^*)\big)^{-1},$$

where the symmetrization matrix drops out because of the structure of the matrices involved.

It appears that (1.24) suffers from a few problems. The first is that the right-hand side contains unknown parameters: not only $\boldsymbol{\gamma}$, but also the very parameters that the left-hand side estimates, through its dependence on $\boldsymbol{V}^*$.

Thus, as before, we have to insert (preliminary) estimators of these. This leads to the following typical estimation procedure: (1) compute the 1-step or 2-step OLS estimate of $\boldsymbol{\gamma}$; (2) use this to compute an estimate of $\boldsymbol{y}^*$ and compute a preliminary estimate of $\boldsymbol{\gamma}^*$ from (1.24) with $\boldsymbol{V}^* = \boldsymbol{I}$; (3) use this to compute an estimate $\tilde{\boldsymbol{V}}$ of $\boldsymbol{V}$ and compute the WLS estimator of $\boldsymbol{\gamma}$ from (1.20); (4) use this to compute an improved estimate of $\boldsymbol{y}^*$ and compute the WLS estimate of $\boldsymbol{\gamma}^*$ from (1.24) with $\boldsymbol{V}^* = 2(\tilde{\boldsymbol{V}} \otimes \tilde{\boldsymbol{V}})$. Variations, e.g., using the estimators of de Leeuw and Kreft [28] as preliminary estimators, are possible, but as it is presented here, it suggests further iterating steps (3) and (4). Indeed, this is typically done and leads to the IGLS algorithm discussed in Section 1.6.2 below.

The second problem with direct application of (1.24) is that it is a computational disaster. The matrix $\boldsymbol{V}^*$ is of order $n^2 \times n^2$, so if $n = 20,000$ as in the application reported below, then we would have to store and invert a 400 million $\times$ 400 million matrix. Fortunately, however, the problem has so much structure that this is not necessary: $\boldsymbol{V}^* = 2(\boldsymbol{V} \otimes \boldsymbol{V})$, which reduces the problem to $n \times n$, but the direct sum form of $\boldsymbol{V}$ reduces this further to $n_j \times n_j$. Then, reductions like the ones used above to arrive at (1.20) as a more convenient version of (1.19) further simplify the computations. Efficient computational procedures are discussed in Goldstein and Rasbash [47].

A variant of (1.24) is obtained by recognizing that the WLS estimator $\underline{\hat{\gamma}}$ that is inserted in the computation of $\boldsymbol{y}^*$ is not equal to $\boldsymbol{\gamma}$, but is an unbiased estimator with variance $(\boldsymbol{U}'\boldsymbol{V}^{-1}\boldsymbol{U})^{-1}$, ignoring variance due to estimation error in the preliminary estimate of $\boldsymbol{V}$. More specifically, by writing

$$\boldsymbol{y} - \boldsymbol{U}\hat{\boldsymbol{\gamma}} = \big[\boldsymbol{I} - \boldsymbol{U}(\boldsymbol{U}'\boldsymbol{V}^{-1}\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{V}^{-1}\big]\boldsymbol{y},$$

it follows that

$$\begin{aligned} &E\big[(\underline{\boldsymbol{y}} - \boldsymbol{U}\underline{\hat{\boldsymbol{\gamma}}})(\underline{\boldsymbol{y}} - \boldsymbol{U}\underline{\hat{\boldsymbol{\gamma}}})'\big] \\ &\quad = \big[\boldsymbol{I} - \boldsymbol{U}(\boldsymbol{U}'\boldsymbol{V}^{-1}\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{V}^{-1}\big]\boldsymbol{V}\big[\boldsymbol{I} - \boldsymbol{U}(\boldsymbol{U}'\boldsymbol{V}^{-1}\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{V}^{-1}\big]' \\ &\quad = \boldsymbol{V} - \boldsymbol{U}(\boldsymbol{U}'\boldsymbol{V}^{-1}\boldsymbol{U})^{-1}\boldsymbol{U}', \end{aligned}$$

or

$$E\big[(\underline{\boldsymbol{y}} - \boldsymbol{U}\underline{\hat{\boldsymbol{\gamma}}})(\underline{\boldsymbol{y}} - \boldsymbol{U}\underline{\hat{\boldsymbol{\gamma}}})' + \boldsymbol{U}(\boldsymbol{U}'\boldsymbol{V}^{-1}\boldsymbol{U})^{-1}\boldsymbol{U}'\big] = \boldsymbol{V}.$$

This suggests replacing (1.23) by

$$\boldsymbol{y}^* \triangleq \operatorname{vec}\big[(\underline{\boldsymbol{y}} - \boldsymbol{U}\underline{\hat{\boldsymbol{\gamma}}})(\underline{\boldsymbol{y}} - \boldsymbol{U}\underline{\hat{\boldsymbol{\gamma}}})' + \boldsymbol{U}(\boldsymbol{U}'\tilde{\boldsymbol{V}}^{-1}\boldsymbol{U})^{-1}\boldsymbol{U}'\big] \tag{1.25}$$

and then proceeding with the estimation process as described above. The term $\boldsymbol{U}(\boldsymbol{U}'\tilde{\boldsymbol{V}}^{-1}\boldsymbol{U})^{-1}\boldsymbol{U}'$ can be viewed as a bias correction. The resulting estimator is again consistent with the same expression for the asymptotic covariance matrix, but is generally less biased in finite samples. The iteration procedure described above with this estimator leads to RIGLS estimators, which are also discussed in Section 1.6.2 below.

1.6.2 Maximum Likelihood

Except for some special cases, explicit closed-form expressions for the maximum likelihood estimators are not available. The loglikelihood function has to be optimized by using some kind of numerical algorithm. This section discusses several of the available algorithms. We can distinguish, on the one hand, generic numerical optimization techniques that can be used for any well-behaved function and, on the other hand, algorithms that are more specific to the problem at hand.

Let $f(\boldsymbol{\theta})$ be a loss function of a parameter vector $\boldsymbol{\theta}$. We want to find the value $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ that minimizes $f(\boldsymbol{\theta})$. Throughout, we assume that $f(\boldsymbol{\theta})$ is well behaved, i.e., that it is continuous and has continuous first and second partial derivatives, is locally Lipschitz, etc. The loss functions for FIML and REML satisfy these and other regularity conditions except in pathological situations where the sample data have no variation or predictor matrices are not of full rank. Thus, we assume these away.

For a short introduction to generic numerical optimization, we refer to Appendix 1.B. The (modified) Newton-Raphson method mentioned there is described for multilevel models by Jennrich and Schluchter [63] and Lindstrom and Bates [73] and it is used in the BMDP5V program [107] for repeated measures models and the nlme package [90] for multilevel analysis in R. The BFGS method is implemented in most general-purpose optimization functions and is used in the MLA program for multilevel analysis [16]. From the discussion in Appendix 1.B, it is clear that we typically need at least first partial derivatives of the loss function, and for Newton-Raphson also the second partial derivatives. We will give their formulas for the FIML and REML loss functions below.

Derivatives of FIML

Computing the partial derivatives of the loglikelihood function with respect to the parameters is a straightforward, albeit tedious, application of (matrix) calculus as developed by, e.g., Magnus and Neudecker [80]. Here we only give the results, the derivations are available from us upon request. Throughout, we will assume that $\boldsymbol{\Omega}$ is parametrized as in (1.3). The first partial derivatives are

$$\frac{\partial \mathcal{L}^F}{\partial \boldsymbol{\gamma}} = -2 \sum_{j=1}^{m} \boldsymbol{Z}_j' \boldsymbol{W}_j^{-1} \boldsymbol{t}_j, \tag{1.26a}$$

$$\frac{\partial \mathcal{L}^F}{\partial \sigma_j^2} = -(n_j - p) \left(\frac{s_j^2 - \sigma_j^2}{(\sigma_j^2)^2} \right) - \operatorname{tr}\left[\boldsymbol{T}_j (\boldsymbol{X}_j' \boldsymbol{X}_j)^{-1} \right], \tag{1.26b}$$

$$\frac{\partial \mathcal{L}^F}{\partial \xi_g} = -\sum_{j=1}^m \text{tr}(\boldsymbol{T}_j \boldsymbol{C}_g), \tag{1.26c}$$

where

$$\begin{aligned}
\boldsymbol{t}_j &\triangleq \boldsymbol{b}_j - \boldsymbol{Z}_j \boldsymbol{\gamma}, \\
\boldsymbol{T}_j &\triangleq \boldsymbol{W}_j^{-1}(\boldsymbol{t}_j \boldsymbol{t}_j' - \boldsymbol{W}_j)\boldsymbol{W}_j^{-1}.
\end{aligned}$$

It is easy to check that the expected values of these partials (when viewed as functions of random variables) are zero, as they should be. It follows immediately from (1.26a) that after convergence (first partials are zero), (1.20) holds. Thus, the FIML estimator of $\boldsymbol{\gamma}$ is a WLS estimator based on the FIML estimates of the variance parameters.

The second partial derivatives with respect to the parameters are

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}^F}{\partial \boldsymbol{\gamma}\, \partial \boldsymbol{\gamma}'} &= 2\sum_{j=1}^m \boldsymbol{Z}_j' \boldsymbol{W}_j^{-1} \boldsymbol{Z}_j, \\
\frac{\partial^2 \mathcal{L}^F}{\partial \boldsymbol{\gamma}\, \partial \sigma_j^2} &= 2\boldsymbol{Z}_j' \boldsymbol{W}_j^{-1}(\boldsymbol{X}_j' \boldsymbol{X}_j)^{-1} \boldsymbol{W}_j^{-1} \boldsymbol{t}_j, \\
\frac{\partial^2 \mathcal{L}^F}{\partial \boldsymbol{\gamma}\, \partial \xi_g} &= 2\sum_{j=1}^m \boldsymbol{Z}_j' \boldsymbol{W}_j^{-1} \boldsymbol{C}_g \boldsymbol{W}_j^{-1} \boldsymbol{t}_j, \\
\frac{\partial^2 \mathcal{L}^F}{\partial \sigma_j^2\, \partial \sigma_j^2} &= (n_j - p)\left(\frac{2s_j^2 - \sigma_j^2}{(\sigma_j^2)^3}\right) + \text{tr}\big[\boldsymbol{\Upsilon}_j (\boldsymbol{X}_j' \boldsymbol{X}_j)^{-1}\big], \\
\frac{\partial^2 \mathcal{L}^F}{\partial \sigma_j^2\, \partial \sigma_k^2} &= 0 \qquad \text{for } k \neq j, \\
\frac{\partial^2 \mathcal{L}^F}{\partial \sigma_j^2\, \partial \xi_g} &= \text{tr}(\boldsymbol{\Upsilon}_j \boldsymbol{C}_g), \\
\frac{\partial^2 \mathcal{L}^F}{\partial \xi_g\, \partial \xi_h} &= \sum_{j=1}^m \text{tr}(\boldsymbol{T}_j \boldsymbol{C}_h \boldsymbol{W}_j^{-1} \boldsymbol{C}_g + \boldsymbol{W}_j^{-1} \boldsymbol{C}_h \boldsymbol{T}_j \boldsymbol{C}_g + \boldsymbol{W}_j^{-1} \boldsymbol{C}_h \boldsymbol{W}_j^{-1} \boldsymbol{C}_g),
\end{aligned}$$

where

$$\boldsymbol{\Upsilon}_j \triangleq \boldsymbol{T}_j (\boldsymbol{X}_j' \boldsymbol{X}_j)^{-1} \boldsymbol{W}_j^{-1} + \boldsymbol{W}_j^{-1} (\boldsymbol{X}_j' \boldsymbol{X}_j)^{-1} \boldsymbol{T}_j + \boldsymbol{W}_j^{-1} (\boldsymbol{X}_j' \boldsymbol{X}_j)^{-1} \boldsymbol{W}_j^{-1}.$$

As mentioned above, often it is assumed that all residual variances are the same: $\sigma_j^2 = \sigma^2$. This leads to fairly trivial changes in these formulas: Every explicit or implicit occurrence of σ_j^2 on the right-hand side is replaced by σ^2, and the derivatives with respect to σ^2 are simply the sums over all groups of the derivatives with respect to σ_j^2 as given here:

$$\frac{\partial \mathcal{L}^F}{\partial \sigma^2} = -\sum_{j=1}^{m} \left\{ (n_j - p) \left(\frac{s_j^2 - \sigma^2}{(\sigma^2)^2} \right) + \operatorname{tr}\left[\boldsymbol{T}_j (\boldsymbol{X}_j' \boldsymbol{X}_j)^{-1} \right] \right\}, \tag{1.27}$$

$$\frac{\partial^2 \mathcal{L}^F}{\partial \boldsymbol{\gamma}\, \partial \sigma^2} = 2 \sum_{j=1}^{m} \boldsymbol{Z}_j' \boldsymbol{W}_j^{-1} (\boldsymbol{X}_j' \boldsymbol{X}_j)^{-1} \boldsymbol{W}_j^{-1} \boldsymbol{t}_j,$$

$$\frac{\partial^2 \mathcal{L}^F}{\partial \sigma^2\, \partial \sigma^2} = \sum_{j=1}^{m} \left\{ (n_j - p) \left(\frac{2 s_j^2 - \sigma^2}{(\sigma^2)^3} \right) + \operatorname{tr}\left[\boldsymbol{\Upsilon}_j (\boldsymbol{X}_j' \boldsymbol{X}_j)^{-1} \right] \right\},$$

$$\frac{\partial^2 \mathcal{L}^F}{\partial \sigma^2\, \partial \xi_g} = \sum_{j=1}^{m} \operatorname{tr}(\boldsymbol{\Upsilon}_j \boldsymbol{C}_g).$$

The derivatives can now be used in a standard numerical optimization algorithm to obtain the FIML estimates.

Derivatives of REML

The first partial derivatives of the REML loss function with respect to the parameters are

$$\frac{\partial \mathcal{L}^R}{\partial \sigma_j^2} = -(n_j - p) \left(\frac{s_j^2 - \sigma_j^2}{(\sigma_j^2)^2} \right) - \operatorname{tr}\left[\boldsymbol{\Delta}_j (\boldsymbol{X}_j' \boldsymbol{X}_j)^{-1} \right], \tag{1.28a}$$

$$\frac{\partial \mathcal{L}^R}{\partial \xi_g} = -\sum_{j=1}^{m} \operatorname{tr}(\boldsymbol{\Delta}_j \boldsymbol{C}_g), \tag{1.28b}$$

where

$$\boldsymbol{\Delta}_j \triangleq \boldsymbol{W}_j^{-1} (\hat{\boldsymbol{t}}_j \hat{\boldsymbol{t}}_j' - \boldsymbol{W}_j + \boldsymbol{Z}_j \boldsymbol{A} \boldsymbol{Z}_j') \boldsymbol{W}_j^{-1},$$

$$\hat{\boldsymbol{t}}_j \triangleq \boldsymbol{b}_j - \boldsymbol{Z}_j \hat{\boldsymbol{\gamma}},$$

$$\hat{\boldsymbol{\gamma}} \triangleq \boldsymbol{A} \sum_{j=1}^{m} \boldsymbol{Z}_j' \boldsymbol{W}_j^{-1} \boldsymbol{b}_j,$$

$$\boldsymbol{A} \triangleq \left(\sum_{j=1}^{m} \boldsymbol{Z}_j' \boldsymbol{W}_j^{-1} \boldsymbol{Z}_j \right)^{-1}.$$

Note that there are no derivatives with respect to $\boldsymbol{\gamma}$, because $\mathcal{L}^R$ is not a function of $\boldsymbol{\gamma}$. We use $\hat{\boldsymbol{\gamma}}$ as a shorthand, but it is not a parameter, it is a function of the data and the variance parameters. Of course, after convergence, this same definition is used to obtain a WLS estimate of $\boldsymbol{\gamma}$, but in deriving statistical properties of the REML estimators, we must treat $\hat{\boldsymbol{\gamma}}$ as a function and not as a mathematical variable.

The second partial derivatives of the REML loss function with respect to the parameters are

$$\begin{aligned}
\frac{\partial^2\mathcal{L}^R}{\partial\sigma_j^2\,\partial\sigma_j^2} &= (n_j-p)\left(\frac{2s_j^2-\sigma_j^2}{(\sigma_j^2)^3}\right)+\text{tr}\big[\boldsymbol{\Theta}_j(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\big]\\
&\quad-2\hat{\boldsymbol{u}}_j'\boldsymbol{A}\hat{\boldsymbol{u}}_j-\text{tr}(\boldsymbol{\Lambda}_j\boldsymbol{A}\boldsymbol{\Lambda}_j\boldsymbol{A}),\\
\frac{\partial^2\mathcal{L}^R}{\partial\sigma_j^2\,\partial\sigma_k^2} &= -2\hat{\boldsymbol{u}}_j'\boldsymbol{A}\hat{\boldsymbol{u}}_k-\text{tr}(\boldsymbol{\Lambda}_j\boldsymbol{A}\boldsymbol{\Lambda}_k\boldsymbol{A})\qquad\text{for } k\neq j,\\
\frac{\partial^2\mathcal{L}^R}{\partial\sigma_j^2\,\partial\xi_g} &= \text{tr}(\boldsymbol{\Theta}_j\boldsymbol{C}_g)-2\hat{\boldsymbol{u}}_j'\boldsymbol{A}\hat{\boldsymbol{\tau}}_g-\text{tr}(\boldsymbol{\Lambda}_j\boldsymbol{A}\boldsymbol{\Xi}_g\boldsymbol{A}),\\
\frac{\partial^2\mathcal{L}^R}{\partial\xi_g\,\partial\xi_h} &= \sum_{j=1}^m\text{tr}(\boldsymbol{\Delta}_j\boldsymbol{C}_h\boldsymbol{W}_j^{-1}\boldsymbol{C}_g+\boldsymbol{W}_j^{-1}\boldsymbol{C}_h\boldsymbol{\Delta}_j\boldsymbol{C}_g+\boldsymbol{W}_j^{-1}\boldsymbol{C}_h\boldsymbol{W}_j^{-1}\boldsymbol{C}_g)\\
&\quad-2\hat{\boldsymbol{\tau}}_h'\boldsymbol{A}\hat{\boldsymbol{\tau}}_g-\text{tr}(\boldsymbol{\Xi}_h\boldsymbol{A}\boldsymbol{\Xi}_g\boldsymbol{A}),
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{\Theta}_j &\triangleq \boldsymbol{\Delta}_j(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{W}_j^{-1}+\boldsymbol{W}_j^{-1}(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{\Delta}_j+\boldsymbol{W}_j^{-1}(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{W}_j^{-1},\\
\boldsymbol{\Lambda}_j &\triangleq \boldsymbol{Z}_j'\boldsymbol{W}_j^{-1}(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{W}_j^{-1}\boldsymbol{Z}_j,\\
\hat{\boldsymbol{u}}_j &\triangleq \boldsymbol{Z}_j'\boldsymbol{W}_j^{-1}(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{W}_j^{-1}\hat{\boldsymbol{t}}_j,\\
\boldsymbol{\Xi}_g &\triangleq \sum_{j=1}^m\boldsymbol{Z}_j'\boldsymbol{W}_j^{-1}\boldsymbol{C}_g\boldsymbol{W}_j^{-1}\boldsymbol{Z}_j,\\
\hat{\boldsymbol{\tau}}_g &\triangleq \sum_{j=1}^m\boldsymbol{Z}_j'\boldsymbol{W}_j^{-1}\boldsymbol{C}_g\boldsymbol{W}_j^{-1}\hat{\boldsymbol{t}}_j.
\end{aligned}$$

When all σ_j^2 are equal, the first partial derivative with respect to σ^2 becomes

$$\frac{\partial\mathcal{L}^R}{\partial\sigma^2}=-\sum_{j=1}^m\left\{(n_j-p)\left(\frac{s_j^2-\sigma^2}{(\sigma^2)^2}\right)+\text{tr}\big[\boldsymbol{\Delta}_j(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\big]\right\}\tag{1.29}$$

and the second partial derivatives involving σ^2 become

$$\begin{aligned}
\frac{\partial^2\mathcal{L}^R}{\partial\sigma^2\,\partial\sigma^2} &= \sum_{j=1}^m\left\{(n_j-p)\left(\frac{2s_j^2-\sigma^2}{(\sigma^2)^3}\right)+\text{tr}\big[\boldsymbol{\Theta}_j(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\big]\right\}\\
&\quad-2\hat{\boldsymbol{u}}'\boldsymbol{A}\hat{\boldsymbol{u}}-\text{tr}(\boldsymbol{\Lambda}\boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{A}),\\
\frac{\partial^2\mathcal{L}^R}{\partial\sigma^2\,\partial\xi_g} &= \sum_{j=1}^m\text{tr}(\boldsymbol{\Theta}_j\boldsymbol{C}_g)-2\hat{\boldsymbol{u}}'\boldsymbol{A}\hat{\boldsymbol{\tau}}_g-\text{tr}(\boldsymbol{\Lambda}\boldsymbol{A}\boldsymbol{\Xi}_g\boldsymbol{A}),
\end{aligned}$$

where

$$\hat{\boldsymbol{u}}\triangleq\sum_{j=1}^m\hat{\boldsymbol{u}}_j\,,$$

$$\boldsymbol{\Lambda} \triangleq \sum_{j=1}^{m} \boldsymbol{\Lambda}_j.$$

Standard Errors

For the standard errors, we need the expectations of the second derivatives instead of the second derivatives themselves. This simplifies the formulas considerably, because many terms have expectation zero and thus drop out. In particular, using $E(\underline{\boldsymbol{t}}_j) = \boldsymbol{\emptyset}$, we obtain

$$E\left(\frac{\partial^2 \underline{\mathcal{L}}^F}{\partial \boldsymbol{\gamma}\, \partial \boldsymbol{\gamma}'}\right) = 2\sum_{j=1}^{m} \boldsymbol{Z}_j' \boldsymbol{W}_j^{-1} \boldsymbol{Z}_j\,,$$

$$E\left(\frac{\partial^2 \underline{\mathcal{L}}^F}{\partial \boldsymbol{\gamma}\, \partial \sigma_j^2}\right) = \boldsymbol{\emptyset},$$

$$E\left(\frac{\partial^2 \underline{\mathcal{L}}^F}{\partial \boldsymbol{\gamma}\, \partial \xi_g}\right) = \boldsymbol{\emptyset}.$$

Hence, the matrix of expectations of the second derivatives of the FIML loss function is a block-diagonal matrix with a diagonal block for the fixed coefficients and a diagonal block for the variance parameters.

For the latter part, we observe that $E(\underline{\boldsymbol{T}}_j) = \boldsymbol{\emptyset}$ implies that

$$E(\underline{\boldsymbol{\Upsilon}}_j) = \boldsymbol{W}_j^{-1}(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{W}_j^{-1}.$$

Consequently,

$$E\left(\frac{\partial^2 \underline{\mathcal{L}}^F}{\partial \sigma_j^2\, \partial \sigma_j^2}\right) = \frac{n_j - p}{(\sigma_j^2)^2} + \operatorname{tr}\big[\boldsymbol{W}_j^{-1}(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{W}_j^{-1}(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\big],$$

$$E\left(\frac{\partial^2 \underline{\mathcal{L}}^F}{\partial \sigma_j^2\, \partial \sigma_k^2}\right) = 0 \qquad \text{for } k \neq j,$$

$$E\left(\frac{\partial^2 \underline{\mathcal{L}}^F}{\partial \sigma_j^2\, \partial \xi_g}\right) = \operatorname{tr}\big[\boldsymbol{W}_j^{-1}(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{W}_j^{-1}\boldsymbol{C}_g\big],$$

$$E\left(\frac{\partial^2 \underline{\mathcal{L}}^F}{\partial \xi_g\, \partial \xi_h}\right) = \sum_{j=1}^{m} \operatorname{tr}(\boldsymbol{W}_j^{-1}\boldsymbol{C}_h\boldsymbol{W}_j^{-1}\boldsymbol{C}_g).$$

When all σ_j^2 are the same, the first three of these are replaced by

$$E\left(\frac{\partial^2 \underline{\mathcal{L}}^F}{\partial \sigma^2\, \partial \sigma^2}\right) = \sum_{j=1}^{m} \left\{ \frac{n_j - p}{(\sigma^2)^2} + \operatorname{tr}\big[\boldsymbol{W}_j^{-1}(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{W}_j^{-1}(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\big] \right\},$$

$$E\left(\frac{\partial^2 \underline{\mathcal{L}}^F}{\partial \sigma^2\, \partial \xi_g}\right) = \sum_{j=1}^{m} \operatorname{tr}\big[\boldsymbol{W}_j^{-1}(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{W}_j^{-1}\boldsymbol{C}_g\big].$$

The information matrix $\mathcal{I}^F$ is defined as

$$\mathcal{I}^F \triangleq E\left(-\frac{\partial^2 \underline{\ell}^F}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\right),$$

where $\underline{\ell}^F$ is the FIML loglikelihood function viewed as a random variable and $\boldsymbol{\theta}$ is the parameter vector. Up till now, we have ignored some constants that do not affect the estimators, but we need to be a little more precise for the standard errors. In fact, $\underline{\mathcal{L}}_j^F = -2(\underline{\ell}_j^F - K_j)$, where K_j is a constant that does not depend on the parameters. Hence, it follows that

$$\mathcal{I}^F = \tfrac{1}{2}\,E\left(\frac{\partial^2 \underline{\mathcal{L}}^F}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\right),$$

so we have to divide the formulas that have just been given by 2. Standard maximum likelihood theory tells us that the standard errors of the estimators are the square roots of the diagonal elements of $(\mathcal{I}^F)^{-1}$. In particular, the submatrix of $\mathcal{I}^F$ corresponding to $\boldsymbol{\gamma}$ is

$$\mathcal{I}^F_{\boldsymbol{\gamma}\boldsymbol{\gamma}} = \sum_{j=1}^m \boldsymbol{Z}_j'\boldsymbol{W}_j^{-1}\boldsymbol{Z}_j.$$

Because of the block-diagonal structure of $\mathcal{I}^F$, it follows that the standard errors of $\hat{\underline{\gamma}}$ are the square roots of the elements of

$$(\mathcal{I}^F_{\boldsymbol{\gamma}\boldsymbol{\gamma}})^{-1} = \left(\sum_{j=1}^m \boldsymbol{Z}_j'\boldsymbol{W}_j^{-1}\boldsymbol{Z}_j\right)^{-1},$$

which corroborates the results obtained earlier for the WLS estimator.

Analogously, for the REML estimators, the expressions are

$$\begin{aligned}
E\left(\frac{\partial^2 \underline{\mathcal{L}}^R}{\partial\sigma_j^2\,\partial\sigma_j^2}\right) &= \frac{n_j - p}{(\sigma_j^2)^2} + \operatorname{tr}\big[\boldsymbol{W}_j^{-1}(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{W}_j^{-1}(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\big] \\
&\quad - \operatorname{tr}(\boldsymbol{\Lambda}_j\boldsymbol{A}\boldsymbol{\Lambda}_j\boldsymbol{A}), \\
E\left(\frac{\partial^2 \underline{\mathcal{L}}^R}{\partial\sigma_j^2\,\partial\sigma_k^2}\right) &= -\operatorname{tr}(\boldsymbol{\Lambda}_j\boldsymbol{A}\boldsymbol{\Lambda}_k\boldsymbol{A}) \qquad \text{for } k \neq j, \\
E\left(\frac{\partial^2 \underline{\mathcal{L}}^R}{\partial\sigma_j^2\,\partial\xi_g}\right) &= \operatorname{tr}\big[\boldsymbol{W}_j^{-1}(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{W}_j^{-1}\boldsymbol{C}_g\big] - \operatorname{tr}(\boldsymbol{\Lambda}_j\boldsymbol{A}\boldsymbol{\Xi}_g\boldsymbol{A}), \\
E\left(\frac{\partial^2 \underline{\mathcal{L}}^R}{\partial\xi_g\,\partial\xi_h}\right) &= \sum_{j=1}^m \operatorname{tr}(\boldsymbol{W}_j^{-1}\boldsymbol{C}_h\boldsymbol{W}_j^{-1}\boldsymbol{C}_g) - \operatorname{tr}(\boldsymbol{\Xi}_h\boldsymbol{A}\boldsymbol{\Xi}_g\boldsymbol{A}).
\end{aligned}$$

When all σ_j^2 are the same, the first three of these are replaced by

$$E\left(\frac{\partial^2 \underline{\mathcal{L}}^R}{\partial\sigma^2\,\partial\sigma^2}\right) = \sum_{j=1}^{m}\left\{\frac{n_j - p}{(\sigma^2)^2} + \operatorname{tr}\left[\boldsymbol{W}_j^{-1}(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{W}_j^{-1}(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\right]\right\} - \operatorname{tr}(\boldsymbol{\Lambda}\boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{A}),$$

$$E\left(\frac{\partial^2 \underline{\mathcal{L}}^R}{\partial\sigma^2\,\partial\xi_g}\right) = \sum_{j=1}^{m}\operatorname{tr}\left[\boldsymbol{W}_j^{-1}(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{W}_j^{-1}\boldsymbol{C}_g\right] - \operatorname{tr}(\boldsymbol{\Lambda}\boldsymbol{A}\boldsymbol{\Xi}_g\boldsymbol{A}).$$

The information matrix $\boldsymbol{\mathcal{I}}^R$ is again obtained by dividing the formulas for the expectations of the second derivatives by 2. Standard errors are the square roots of the diagonal elements of the inverse of the information matrix.

As indicated above, after convergence, we use the expression for $\hat{\boldsymbol{\gamma}}$ used in the expressions for the REML derivatives as an estimator of $\boldsymbol{\gamma}$. It is immediately clear that this is a WLS estimator with $\boldsymbol{W}_j$ based on the REML estimators for the variance parameters. Hence, the standard error formulas given for WLS above apply directly to this estimator.

Scoring

We have seen above that expressions for the second derivatives of the ML loss functions are rather unwieldy, whereas the expressions for their expectations are much simpler. In fact, because the asymptotic covariance matrix of the estimators is a positive constant times the inverse of the matrix of expected second derivatives, the matrix of expected second derivatives must be a positive definite matrix. Furthermore, in large samples, the exact second derivatives should be close to the expected second derivatives. Combining these statistical observations with the general theory of numerical optimization suggests that a convenient alternative to the Newton-Raphson algorithm would be to replace the Hessian by its expectation. Because the expected Hessian is guaranteed to be positive definite, this does not need to be checked and modifications of it are not necessary. Thus, an easier expression is used, which is computationally less demanding, and the block-diagonality of the expected Hessian reduces the computational burden in computing the inverse as well.

The resulting algorithm, which is specific to loglikelihood functions (but certainly not to multilevel models), is called *Method of Scoring*, *Fisher scoring*, or simply *Scoring*. It was proposed for multilevel models by Longford [74] and implemented in the VARCL program [75]. It tends to be very fast and stable.

Iteratively Reweighted Least Squares

In (1.20), we have seen a simple, yet statistically efficient estimator of the fixed coefficients $\boldsymbol{\gamma}$, given knowledge of the variance parameters. In practice,

this means that consistent estimators of the variance parameters are plugged in. Conversely, in (1.24), combined with either (1.23) or (1.25), we have given a (conceptually) simple and statistically efficient estimator of the variance parameters $\boldsymbol{\gamma}^*$, given $\boldsymbol{\gamma}$ and a preliminary estimate of the variance parameters. As noted there, this suggests an iterative algorithm, in which these two steps are alternated.

This algorithm was introduced for multilevel models by Goldstein [44] using (1.23) to compute $\boldsymbol{y}^*$ and by Goldstein [45] using (1.25) to compute $\boldsymbol{y}^*$. In the former case, the algorithm is called *iterative generalized least squares* (IGLS), whereas in the latter, it is called *restricted iterative generalized least squares* (RIGLS). Similar procedures, also known as *iterative reweighted least squares* (IRLS), are used in many branches of statistics. For example, the standard estimation method for generalized linear models is IRLS [82] and it can be used to compute estimators based on "robust" loss functions, which are less sensitive to outliers [48]. An overview, relating IGLS to various numerical optimization algorithms, is given by del Pino [32]. From these sources, it is known that IGLS produces maximum likelihood estimators.

The equivalence of IGLS to FIML was shown explicitly for the multilevel model by Goldstein [44]. Goldstein [45] showed that RIGLS gives REML estimators. Paralleling his proofs, we can see here, as we have noted above, that setting (1.26a) to zero is equivalent to the IGLS/RIGLS condition (1.20). Furthermore, it is easy to show that (1.24) combined with (1.23) and (1.20) implies that (1.26b) and (1.26c) are zero. Thus, after convergence of the IGLS algorithm, the first partial derivatives of the FIML loglikelihood are zero and, thus (assuming regularity), the IGLS estimates must be equal to the FIML estimates. Analogously, it is equally easy to show that (1.24) combined with (1.25) and (1.20) implies that (1.28a) and (1.28b) are zero and, thus, that after convergence, the RIGLS estimates are equal to the REML estimates.

EM Algorithm

The *EM algorithm* is an iterative method for optimizing functions of the form $f(\boldsymbol{\theta}) = \log \int g(\boldsymbol{\theta}, \boldsymbol{z}) \, \mathrm{d}\boldsymbol{z}$ with respect to $\boldsymbol{\theta}$. It was presented in its full generality by Dempster et al. [33]. Typically, $f(\boldsymbol{\theta})$ is a loglikelihood function and $\log g(\boldsymbol{\theta}, \boldsymbol{z})$ the *complete-data loglikelihood* function, i.e., the loglikelihood function that would have been obtained if the realization of the random variables $\underline{\boldsymbol{z}}$ would have been observed. Thus, both are also implicitly functions of the observed data $\boldsymbol{y}$. Maximization of $f(\boldsymbol{\theta})$ proceeds by iteratively maximizing the expectation of the complete-data loglikelihood. That is, in each iteration, the function

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(i)}) \stackrel{\Delta}{=} E\left[\log g(\boldsymbol{\theta}, \underline{\boldsymbol{z}}) \mid \boldsymbol{y}, \boldsymbol{\theta}^{(i)}\right]$$

is maximized, where the expectation is taken over the conditional distribution of $\underline{\boldsymbol{z}}$ given the observed data $\boldsymbol{y}$ and the value $\boldsymbol{\theta}^{(i)}$ of the parameter vector after the previous iteration. Appendix 1.D explains in more detail why this works.

For the multilevel model, $\underline{\boldsymbol{z}}$ consists of the random effects $\{\underline{\boldsymbol{\delta}}_j\}$, and $\boldsymbol{\theta}$ and $\boldsymbol{y}$ have their usual meaning. As derived in Appendix 1.D, when applied to the FIML loglikelihood, this means that in the expectation step, the following quantities are computed:

$$\begin{aligned}\boldsymbol{\mu}_j^{(i)} &= \boldsymbol{\Omega}\boldsymbol{W}_j^{-1}(\boldsymbol{b}_j - \boldsymbol{Z}_j\boldsymbol{\gamma}),\\ \boldsymbol{\Sigma}_j^{(i)} &= \sigma_j^2\boldsymbol{\Omega}\boldsymbol{W}_j^{-1}(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1},\end{aligned}$$

where the right-hand sides are evaluated in $\boldsymbol{\theta}^{(i)}$. If $\boldsymbol{\Omega}$ is completely free (apart from the requirements of symmetry and positive definiteness, of course), the maximization step leads to the updates

$$\begin{aligned}\boldsymbol{\Omega}^{(i+1)} &= \frac{1}{m}\sum_{j=1}^{m}(\boldsymbol{\Sigma}_j^{(i)} + \boldsymbol{\mu}_j^{(i)}\boldsymbol{\mu}_j^{(i)\prime}),\\ \boldsymbol{\gamma}^{(i+1)} &= \left(\sum_{j=1}^{m}\boldsymbol{Z}_j'\boldsymbol{X}_j'\boldsymbol{X}_j\boldsymbol{Z}_j\right)^{-1}\sum_{j=1}^{m}\boldsymbol{Z}_j'\boldsymbol{X}_j'\boldsymbol{X}_j(\boldsymbol{b}_j - \boldsymbol{\mu}_j^{(i)}),\\ (\sigma_j^2)^{(i+1)} &= \frac{1}{n_j}\left[(n_j - p)s_j^2 + \operatorname{tr}(\boldsymbol{X}_j'\boldsymbol{X}_j\boldsymbol{\Lambda}_j^{(i)})\right],\end{aligned}$$

or, instead of the latter,

$$(\sigma^2)^{(i+1)} = \frac{1}{n}\sum_{j=1}^{m}\left[(n_j - p)s_j^2 + \operatorname{tr}(\boldsymbol{X}_j'\boldsymbol{X}_j\boldsymbol{\Lambda}_j^{(i)})\right],$$

where

$$\boldsymbol{\Lambda}_j^{(i)} \stackrel{\Delta}{=} \boldsymbol{\Sigma}_j^{(i)} + (\boldsymbol{b}_j - \boldsymbol{\mu}_j^{(i)} - \boldsymbol{Z}_j\boldsymbol{\gamma}^{(i+1)})(\boldsymbol{b}_j - \boldsymbol{\mu}_j^{(i)} - \boldsymbol{Z}_j\boldsymbol{\gamma}^{(i+1)})'.$$

If $\boldsymbol{\Omega}$ is restricted, typically by (1.3) with $G < p(p+1)/2$ parameters, the update of the variance parameters $\boldsymbol{\xi}$ is a bit more complicated; see Appendix 1.D.

A great advantage of the EM algorithm is that the loglikelihood is improved in each iteration, i.e., the algorithm is monotonic. Furthermore, the computations in each iteration are often very simple, much simpler than with other numerical optimization algorithms. Another strength of the EM algorithm is that it is able to deal with missing data in a very natural way (see Chapter 10). A drawback of EM is that it tends to converge very slowly. Formally, it converges linearly, whereas, for example, Newton-Raphson converges quadratically when in the neighborhood of the optimum. On the other hand, when far from the optimum, the EM algorithm shows more stable convergence in the direction of the optimum. For this reason, the nlme package [90] uses

EM for the initial iterations and switches to Newton-Raphson later on in the algorithm. An incomplete list of other multilevel packages that use EM, either as an option or for specific tasks, is BMDP-5V [107], MLA [16], and especially HLM [102], which popularized the algorithm for multilevel analysis. The EM algorithm is described for multilevel analysis and especially its special case of repeated measures models in Dempster et al. [34], Laird and Ware [70], Jennrich and Schluchter [63], Laird et al. [69], Lindstrom and Bates [73], and Raudenbush and Bryk [101, Chapter 14].

Further Numerical and Computational Issues

As we have seen, most formulas for computing estimates for multilevel models can be expressed in different ways. Some of these are clearly computationally inefficient, whereas others use the structure of the problem in better ways. This pertains to usage of memory, sizes of inverses needed, and other ways to compute the same expressions. Given the sizes of typical multilevel datasets and the ways in which computations can be done inefficiently, implementing an estimator for a multilevel model for general use needs considerable fine-tuning.

In many cases, we have presented results using $\boldsymbol{Z}_j$, $\boldsymbol{W}_j$, $\boldsymbol{b}_j$, and a few other matrices and vectors. These are of smaller sizes than $\boldsymbol{U}_j$, $\boldsymbol{V}_j$, and $\boldsymbol{y}_j$, so that this already improves the computations considerably. Longford [74] gives further computational formulas, such that the amount of storage needed is further reduced (but dimensions of inverses do not become smaller).

However, our formulas still use expressions like $\boldsymbol{b}_j = (\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j'\boldsymbol{y}_j$. Actually computing an estimator in this way is generally considered undesirable, because it exacerbates any numerical problems that may exist. A good way to compute a least squares estimator is to use the QR decomposition. Pinheiro and Bates [89] discuss these issues at length and present detailed analyses in which the multilevel loglikelihood is transformed in a way that makes computations fast, numerically stable, and memory efficient. We do not present these here, but recommend their book to interested readers.

1.6.3 Robust Covariance Matrix Estimation

We have seen above that the two-step OLS estimator of $\boldsymbol{\gamma}$ is

$$\hat{\underline{\gamma}} = \left(\sum_{j=1}^m \boldsymbol{Z}_j'\boldsymbol{Z}_j\right)^{-1}\sum_{j=1}^m \boldsymbol{Z}_j'\underline{\boldsymbol{b}}_j = \boldsymbol{A}\sum_{j=1}^m \boldsymbol{Z}_j'\underline{\boldsymbol{b}}_j,$$

with $\boldsymbol{A}$ implicitly defined. Its covariance matrix is

$$\boldsymbol{C} \triangleq \mathrm{Cov}(\hat{\underline{\gamma}}) = \boldsymbol{A}\left(\sum_{j=1}^m \boldsymbol{Z}_j'\,\mathrm{Cov}(\underline{\boldsymbol{b}}_j)\boldsymbol{Z}_j\right)\boldsymbol{A}.$$

If $m \to \infty$, $\hat{\underline{\gamma}}$ is a consistent estimator of $\boldsymbol{\gamma}$, and instead of using the model-based estimator of $\boldsymbol{C}$ presented earlier, $\boldsymbol{C}$ can be straightforwardly estimated by the cluster-robust covariance matrix [e.g., 98]

$$\hat{\underline{\boldsymbol{C}}}_{\text{cr}} = \boldsymbol{A}\left(\sum_{j=1}^{m} \boldsymbol{Z}_j'\hat{\underline{\boldsymbol{t}}}_j\hat{\underline{\boldsymbol{t}}}_j'\boldsymbol{Z}_j\right)\boldsymbol{A},$$

where $\hat{\underline{\boldsymbol{t}}}_j = \underline{\boldsymbol{b}}_j - \boldsymbol{Z}_j\hat{\underline{\gamma}}$. When m is large, this is an accurate estimator, but in moderately large samples, it tends to be biased because the variability in estimation of $\boldsymbol{\gamma}$ is not taken into account. That is, the difference between $\hat{\underline{\boldsymbol{t}}}_j$ and $\underline{\boldsymbol{t}}_j \stackrel{\Delta}{=} \underline{\boldsymbol{b}}_j - \boldsymbol{Z}_j\boldsymbol{\gamma}$ is ignored. Inspired by similar problems with the (Eicker-Huber-)White heteroskedasticity-consistent covariance matrix, and fairly successful corrections thereof [25, pp. 552–556], corrections to the cluster-robust covariance matrix can be computed, which take the form of multiplication by a certain factor, e.g.,

$$\frac{m}{m-1}\,\frac{n-1}{n-r},$$

where n is total sample size and r is the number of elements of $\boldsymbol{\gamma}$. Cameron and Trivedi [17, p. 834] mention this correction in the context of the one-step OLS estimator.

Analogously, abusing the same notation for different estimators, the one-step OLS estimator is

$$\hat{\underline{\gamma}} = \left(\sum_{j=1}^{m} \boldsymbol{U}_j'\boldsymbol{U}_j\right)^{-1} \sum_{j=1}^{m} \boldsymbol{U}_j'\underline{\boldsymbol{y}}_j = \boldsymbol{A}\sum_{j=1}^{m} \boldsymbol{U}_j'\underline{\boldsymbol{y}}_j.$$

Thus, we can estimate its covariance matrix by the cluster-robust covariance estimator

$$\hat{\underline{\boldsymbol{C}}}_{\text{cr}} = \boldsymbol{A}\left(\sum_{j=1}^{m} \boldsymbol{U}_j'\hat{\underline{\boldsymbol{r}}}_j\hat{\underline{\boldsymbol{r}}}_j'\boldsymbol{U}_j\right)\boldsymbol{A},$$

[e.g., 126, p. 152], where $\hat{\underline{\boldsymbol{r}}}_j = \underline{\boldsymbol{y}}_j - \boldsymbol{U}_j\hat{\underline{\gamma}}$. As observed above, the one-step OLS estimator can also be written as

$$\hat{\underline{\gamma}} = \boldsymbol{A}\sum_{j=1}^{m} \boldsymbol{Z}_j'\boldsymbol{X}_j'\boldsymbol{X}_j\underline{\boldsymbol{b}}_j,$$

where $\boldsymbol{A}$ is now written as

$$\boldsymbol{A} = \left(\sum_{j=1}^{m} \boldsymbol{Z}_j'\boldsymbol{X}_j'\boldsymbol{X}_j\boldsymbol{Z}_j\right)^{-1}.$$

Hence, the cluster-robust covariance estimator can be rewritten as

$$\underline{\hat{\boldsymbol{C}}}_{\text{cr}} = \boldsymbol{A}\left(\sum_{j=1}^{m}\boldsymbol{Z}_j'\boldsymbol{X}_j'\boldsymbol{X}_j\underline{\hat{\boldsymbol{t}}}_j\underline{\hat{\boldsymbol{t}}}_j'\boldsymbol{X}_j'\boldsymbol{X}_j\boldsymbol{Z}_j\right)\boldsymbol{A},$$

where it is now natural to use the one-step estimator of the coefficient vector $\boldsymbol{\gamma}$ in the definition of $\underline{\hat{\boldsymbol{t}}}_j$.

In the same way, a straightforward cluster-robust covariance matrix of the WLS estimator $\underline{\hat{\boldsymbol{\gamma}}}$ is found to be

$$\underline{\hat{\boldsymbol{C}}}_{\text{cr}} = \underline{\hat{\boldsymbol{A}}}\left(\sum_{j=1}^{m}\boldsymbol{U}_j'\underline{\hat{\boldsymbol{V}}}_j^{-1}\underline{\hat{\boldsymbol{r}}}_j\underline{\hat{\boldsymbol{r}}}_j'\underline{\hat{\boldsymbol{V}}}_j^{-1}\boldsymbol{U}_j\right)\underline{\hat{\boldsymbol{A}}},$$

where now the WLS estimator of $\boldsymbol{\gamma}$ is used in the definition of $\underline{\hat{\boldsymbol{r}}}_j$,

$$\begin{aligned}\underline{\hat{\boldsymbol{V}}}_j &= \boldsymbol{X}_j\underline{\hat{\boldsymbol{\Omega}}}\boldsymbol{X}_j' + \underline{\hat{\sigma}}_j^2\boldsymbol{I}_{n_j},\\ \underline{\hat{\boldsymbol{A}}} &= \left(\sum_{j=1}^{m}\boldsymbol{U}_j'\underline{\hat{\boldsymbol{V}}}_j^{-1}\boldsymbol{U}_j\right)^{-1},\end{aligned}$$

or, equivalently,

$$\underline{\hat{\boldsymbol{C}}}_{\text{cr}} = \underline{\hat{\boldsymbol{A}}}\left(\sum_{j=1}^{m}\boldsymbol{Z}_j'\underline{\hat{\boldsymbol{W}}}_j^{-1}\underline{\hat{\boldsymbol{t}}}_j\underline{\hat{\boldsymbol{t}}}_j'\underline{\hat{\boldsymbol{W}}}_j^{-1}\boldsymbol{Z}_j\right)\underline{\hat{\boldsymbol{A}}},$$

with

$$\begin{aligned}\underline{\hat{\boldsymbol{W}}}_j &= \underline{\hat{\boldsymbol{\Omega}}} + \underline{\hat{\sigma}}_j^2(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1},\\ \underline{\hat{\boldsymbol{A}}} &= \left(\sum_{j=1}^{m}\boldsymbol{Z}_j'\underline{\hat{\boldsymbol{W}}}_j^{-1}\boldsymbol{Z}_j\right)^{-1},\end{aligned}$$

and the WLS estimator of $\boldsymbol{\gamma}$ is used in the definition of $\underline{\hat{\boldsymbol{t}}}_j$. Note that for the asymptotic results, it does not matter which estimators of $\boldsymbol{\Omega}$ and σ_j^2 are used, as long as they are consistent. Of course, in finite samples, it does matter and we would expect that more precise estimators of $\boldsymbol{\Omega}$ and σ_j^2 result in better estimators of $\boldsymbol{\gamma}$ and $\boldsymbol{C}$.

Robust Covariance Matrices for ML Estimators

A robust covariance estimator for the FIML estimator of $\boldsymbol{\gamma}$ is immediately obtained from the one for the WLS estimator given above. The same applies to the two-step ML ("REML") estimator obtained as a WLS estimator that uses the REML estimates of the variance parameters in computing the weight matrix.

It is also possible to compute a robust covariance matrix for the variance parameters. However, because no closed-form expression for the estimators of the variance parameters exists, this requires a bit more asymptotic statistical theory. The basic idea starts from the first-order condition for ML estimators

$$\sum_{j=1}^{m} \frac{\partial \underline{\mathcal{L}}_j}{\partial \boldsymbol{\theta}}(\hat{\underline{\boldsymbol{\theta}}}) = \boldsymbol{\emptyset}.$$

Then a first-order Taylor series expansion of this, around the true value $\boldsymbol{\theta}_0$, is taken, giving

$$\sum_{j=1}^{m} \left\{ \frac{\partial \underline{\mathcal{L}}_j}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_0) + \frac{\partial^2 \underline{\mathcal{L}}_j}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}(\boldsymbol{\theta}_0)\ (\hat{\underline{\boldsymbol{\theta}}} - \boldsymbol{\theta}_0) + O_p \|\hat{\underline{\boldsymbol{\theta}}} - \boldsymbol{\theta}_0\|^2 \right\} = \boldsymbol{\emptyset}.$$

Under suitable regularity conditions, a form of the central limit theorem implies that

$$\frac{1}{\sqrt{m}} \sum_{j=1}^{m} \frac{\partial \underline{\mathcal{L}}_j}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_0) \stackrel{\mathcal{L}}{\Longrightarrow} \mathcal{N}(\boldsymbol{\emptyset}, \boldsymbol{\Psi})$$

from some finite positive definite matrix $\boldsymbol{\Psi}$, and a form of the law of large numbers implies that

$$\frac{1}{m} \sum_{j=1}^{m} \frac{\partial^2 \underline{\mathcal{L}}_j}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}(\boldsymbol{\theta}_0) \stackrel{\mathcal{P}}{\Longrightarrow} \boldsymbol{\mathcal{H}}$$

for some finite positive definite matrix $\boldsymbol{\mathcal{H}}$. Combining results, we obtain

$$\sqrt{m}(\hat{\underline{\boldsymbol{\theta}}} - \boldsymbol{\theta}_0) = -\boldsymbol{\mathcal{H}}^{-1} \frac{1}{\sqrt{m}} \sum_{j=1}^{m} \frac{\partial \underline{\mathcal{L}}_j}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_0) + o_p(1) \stackrel{\mathcal{L}}{\Longrightarrow} \mathcal{N}(\boldsymbol{\emptyset}, \boldsymbol{\mathcal{H}}^{-1} \boldsymbol{\Psi} \boldsymbol{\mathcal{H}}^{-1}).$$

Obviously, consistent estimators of $\boldsymbol{\mathcal{H}}$ and $\boldsymbol{\Psi}$ are

$$\hat{\underline{\boldsymbol{\mathcal{H}}}} = \frac{1}{m} \sum_{j=1}^{m} \frac{\partial^2 \underline{\mathcal{L}}_j}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}(\hat{\underline{\boldsymbol{\theta}}}),$$

$$\hat{\underline{\boldsymbol{\Psi}}} = \frac{1}{m} \sum_{j=1}^{m} \frac{\partial \underline{\mathcal{L}}_j}{\partial \boldsymbol{\theta}}(\hat{\underline{\boldsymbol{\theta}}})\ \frac{\partial \underline{\mathcal{L}}_j}{\partial \boldsymbol{\theta}'}(\hat{\underline{\boldsymbol{\theta}}}).$$

For computing a robust covariance matrix for $\hat{\underline{\boldsymbol{\theta}}}$, all factors of m drop out and we obtain

$$\hat{\underline{\boldsymbol{C}}}_{\text{cr}} = \left(\sum_{j=1}^{m} \frac{\partial^2 \underline{\mathcal{L}}_j}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}(\hat{\underline{\boldsymbol{\theta}}}) \right)^{-1} \left(\sum_{j=1}^{m} \frac{\partial \underline{\mathcal{L}}_j}{\partial \boldsymbol{\theta}}(\hat{\underline{\boldsymbol{\theta}}})\ \frac{\partial \underline{\mathcal{L}}_j}{\partial \boldsymbol{\theta}'}(\hat{\underline{\boldsymbol{\theta}}}) \right) \left(\sum_{j=1}^{m} \frac{\partial^2 \underline{\mathcal{L}}_j}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}(\hat{\underline{\boldsymbol{\theta}}}) \right)^{-1}. \quad (1.30)$$

The theory underlying the robust covariance matrices for ML estimators in a multilevel model is derived in detail, with appropriate regularity conditions, in Verbeke and Lesaffre [120, 121].

From (1.26), (1.27), (1.28), and (1.29), it follows that this theory should work for the FIML and REML estimators of $\boldsymbol{\gamma}$ and ξ_g, and for the corresponding estimators of σ^2 if all residual variances are assumed to be the same. However, if separate residual variances σ_j^2 are estimated, the corresponding first-order conditions do not satisfy the central limit theorem as presented here, because they have only one term. In that case, assuming that $n_j \rightarrow \infty$, it is still possible to derive some kind of robust variance estimators for the variance estimators $\hat{\underline{\sigma}}_j^2$, using within-groups asymptotics along the lines of Browne [12], but this tends to require large within-group sample sizes, so this may not work well in practice.

Note that when all the model assumptions are met, we have the well-known result (correcting for our scaling of the loglikelihood)

$$\tfrac{1}{2}\boldsymbol{\mathcal{H}} = \tfrac{1}{4}\boldsymbol{\Psi} = \lim_{m\rightarrow\infty}\frac{1}{m}\boldsymbol{\mathcal{I}},$$

which leads to the standard (model-based) covariance matrix presented earlier.

Robust Versus Model-Based Covariance Matrices

With a few exceptions, the model-based covariance matrices are only correct if the complete model is correctly specified ("true"). The robust covariance matrices are consistent under a wider range of assumptions, including fairly general forms of misspecification of the random part of the model, such as intragroup dependence and heteroskedasticity. So if the main interest of the analyses is the fixed part of the model (i.e., $\boldsymbol{\gamma}$), a cluster-robust covariance matrix may be preferred.

On the other hand, if the random part of the model is the main focus of interest, i.e., modeling/explaining between-group variation is important, then an estimator of the covariance matrix of the fixed part that is robust to misspecification of the random part is only of secondary interest. If the random part is (severely) misspecified, the primary aim of the analysis is not met. This is even more salient for robust covariance matrices of the variance parameters themselves. If the model is misspecified, it is generally unclear what is estimated, and thus it is questionable whether a robust covariance matrix is of any use [39].

There is, however, a leading example where the random part is misspecified, but the estimators are still consistent estimators of meaningful parameters. This is the case when the model is correctly specified, except for the distribution of the random variables. If these are nonnormally distributed, the model-based covariance matrices for the estimators of $\boldsymbol{\gamma}$ are still correct, but standard model-based covariance matrices of the variance parameters are incorrect. But $\boldsymbol{\Omega}$ and σ^2 are still meaningful parameters and their estimators

are consistent. So then using a robust covariance matrix is clearly useful [120, 121].

The robust covariance matrices are typically far less precise if the model is (approximately) correctly specified and the sample size is small to moderate. Therefore, in not-too-large samples, the model-based covariance matrices will typically be preferred if the analyst believes that the random part of the model is reasonably well specified. Maas and Hox [78] performed a simulation study to investigate these issues for REML estimators and concluded that the model-based standard errors of the estimator of σ^2 performed well under nonnormality, while the robust standard errors are often too large. However, both model-based and robust standard errors of level-2 variance parameters did not perform very well at small sample size, although the robust ones were clearly better than the model-based ones. They conclude that at least 100 groups are needed for reliable robust standard errors. As a general strategy, they recommend comparing the robust standard errors with the model-based ones to diagnose possible misspecification of the model.

An alternative way for robust statistical inference under possible misspecification is to use resampling methods. Moreover, the bootstrap in particular has the additional potential advantage that it can generate asymmetric confidence intervals, thereby reflecting nonnormal finite-sample distributions of especially the level-2 variance parameters. However, confidence intervals based on resampling methods tend to perform less than satisfactory as well with small or moderate level-2 sample sizes. See Chapter 11 for a detailed description of resampling methods for multilevel models and their empirical properties.

1.7 Software

We will be brief about software here, if only because details about software are likely to be quickly outdated. An overview of the history of the development of software for multilevel analysis, and the state of affairs ca. 2000 is given in de Leeuw and Kreft [30]. The overview is still broadly valid, except that the details have changed and there are some additions.

As mentioned earlier in this chapter, the software packages have largely been developed by the same authors who pioneered the development of multilevel analysis as a statistical method and who have written successful textbooks about multilevel modeling. And, for that matter, are contributors to this Handbook.

Two software packages dominate the market for dedicated multilevel analysis software. These are HLM [102] and MLwiN [97]. These packages offer a broad range of linear and nonlinear specifications of multilevel models and have user-friendly graphical user interfaces. There are some differences in the

algorithms used, but these are not particularly interesting for the average user. There are also some differences in the more advanced options or less frequently used model specifications, so users with specific desires may prefer one over the other for this reason.

Originally, VARCL [75] was also one of the major packages, but development of this package has been terminated. There are many packages that focus on more specific multilevel models, options, or other aspects. These tend to be research software, with fewer options and less user-friendly interfaces, and development of these progresses faster if the authors are working on new directions in their research that requires additions to the programs. Examples of these are MLA [16], which focuses on resampling methods (see Chapter 11) and PINT [10], which focuses on power calculations (see Chapter 4). The MIXFOO suite [55, 56, 57, etc.] also belongs in this category, although taken as a whole, it is a fairly comprehensive multilevel package.

The BUGS program and its variants, most notably WinBUGS [113], are programs for Bayesian data analysis. They offer extensive possibilities for Bayesian multilevel analysis and are particularly useful for estimating nonlinear multilevel models. See also Chapter 2.

Many general-purpose (or almost-all-encompassing) statistical packages now have multilevel options as well. Important examples are SAS® [106], which has PROC MIXED and PROC NLMIXED, SPSS® [114], which has MIXED and several other procedures that can be used for multilevel analyses, Stata® [115], which has many "survey", "cluster", and "panel" programs and options, and the extensive `gllamm` program [95], and R [93], for which the lme4 and nlme packages are available [7, 90].

A relatively recent development is the incorporation of multilevel facilities in programs for structural equation modeling, such as LISREL [35, 64], EQS [8], and Mplus [85]. The possibilities of these programs are somewhat different from the standard multilevel programs. They often have less extensive options for estimating nonlinear models and models with three or more levels, but are better equipped for estimating multivariate models and models with latent variables and measurement errors, i.e., multilevel structural equation models (see Chapter 12). Thus, they complement traditional multilevel packages.

Throughout this Handbook, other software packages (perhaps less well known or more specialized) are mentioned where appropriate and useful.

1.8 Sampling Weights

Surveys are often nonrepresentative of the population of interest, in the sense that persons (or, more generally, units) with certain characteristics are more prevalent in the data than in the population. There are essentially two reasons for this: deliberate oversampling of certain groups and different nonresponse

rates. An example of the former is the oversampling of relatively small groups, like minorities, to obtain more reliable information about these groups. An example of the latter is the tendency to obtain an overrepresentation of women in a study that was designed to be neutral, which may happen because women tend to be more often at home than men.

Agencies that collect such surveys typically provide *sampling weights* with the data set. The idea is that applying these sampling weights in the analysis corrects for the nonrepresentativeness of the data by giving underrepresented groups more weight and overrepresented groups less weight. For example, assume that we are interested in the mean height of adults in a country of interest. Assume further that we have a sample of 1000 adults, 600 of which are women, whereas in the population 50% of adults is female. Height is expected to be related to sex, so if we simply computed the sample average, we would likely obtain an underestimate of our parameter of interest. However, if we give women a weight of $w_i = 5/6$ and men a weight of $w_i = 5/4$, then the weighted average

$$\begin{aligned} \bar{h}_w &\triangleq \frac{\sum_{i=1}^{1000} w_i h_i}{\sum_{i=1}^{1000} w_i} \\ &= \frac{600 \cdot (5/6) \cdot \bar{h}_f + 400 \cdot (5/4) \cdot \bar{h}_m}{600 \cdot (5/6) + 400 \cdot (5/4)} \\ &= 0.5\bar{h}_f + 0.5\bar{h}_m \end{aligned} \tag{1.31}$$

is clearly (the realization of) an unbiased estimator of average height in the population, where h_i is the height of the i-th observation in the sample and $\bar{h}_f$ and $\bar{h}_m$ are the average heights of females and males in the sample, respectively. (Note that apparently some software packages define weights as the reciprocals of the definition we use here, so check your manuals.)

For regression models, there is some discussion in the literature about whether weights should be applied, even if the sample is nonrepresentative and weights are available. In fact, if the standard regression model $\underline{y}_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \underline{\epsilon}_i$, with $\underline{\epsilon}_i$ i.i.d., holds and the nonrepresentativeness is possibly related to $\boldsymbol{x}$ but not to $\underline{\epsilon}$, then OLS is still the most efficient estimator, and all statistical inference is correct. However, in many circumstances, it is quite likely that the error term represents the influence of a large number of variables that each have a fairly small effect, most of which are unknown and/or unobserved, but some of which may be somehow related to the probabilities of being included in the sample. In such cases, OLS would be biased, whereas a weighted analysis would still give an unbiased estimator.

An important special case where a weighted analysis gives simple consistent estimators and an unweighted analysis does not is in the analysis of so-called *choice-based samples* or, more generally, *endogenously stratified* samples. In this case, samples are drawn from strata defined by the dependent

variable. An example is a sample consisting of 500 bus passengers sampled on board bus lines and 500 car drivers sampled along the road, and the dependent variable is mode choice. Another important example is a medical study in which a sample of people having a rare disease is drawn from hospital records and a similar-sized sample of people not having the disease is drawn from the general public, and the dependent variable in the study is whether or not one has the disease.

These issues are extensively discussed in Cameron and Trivedi [17, pp. 817–829] and Wooldridge [124, 125, 127], who also give detailed derivations and explanations, showing why unweighted analyses are sometimes inconsistent and under different circumstances consistent and efficient. For the remainder of this section, we assume that a weighted analysis is desired.

For multilevel analysis, an additional complication is how to deal with units at different levels. To continue our example, assume that we have a two-level sample, where level-1 is individuals and level-2 is counties. Perhaps heights are correlated within counties because of environmental factors, different socio-economic composition, different ethnic composition or more specifically family relations, and therefore a multilevel approach is desired, but still females are overrepresented. Furthermore, let us assume that we know the population percentages of males and females in each county (not necessarily 50%). Then a straightforward adaptation of (1.31) gives an estimate of the within-county mean height:

$$\bar{h}_{wj} \triangleq \frac{\sum_{i=1}^{n_j} w_{i|j} h_{ij}}{\sum_{i=1}^{n_j} w_{i|j}}$$

in obvious notation. If each county had the same population size (or height was unrelated to population size) and the sample of counties is representative of all counties in whatever way this is defined, a simple average of the county averages gives an unbiased estimate of the parameter of interest. More generally, however, we also have a county weight w_j, and the overall weighted mean is computed as

$$\bar{h}_{w\cdot} \triangleq \frac{\sum_{j=1}^{m} w_j \bar{h}_{wj}}{\sum_{j=1}^{m} w_j}.$$

Determining the value of w_j depends on the sampling scheme and the resulting representativeness at the county level. For example, if the counties are a simple random sample of all counties in the country, then counties with small population size are overrepresented given that we are interested in the mean height of individuals. It is easy to see then that w_j should be proportional to county population size N_j. Often, however, sampling at county level is done proportional to size, so that w_j should be the same for each county.

When a survey data set is given, it typically contains an individual weight w_{ij} and the clusters are defined by the researcher. Then the multilevel weights can be computed as

$$w_j \stackrel{\Delta}{=} \sum_{i=1}^{n_j} w_{ij},$$
$$w_{i|j} \stackrel{\Delta}{=} w_{ij}/w_j.$$

See, however, Potthoff et al. [92], Pfeffermann et al. [88], Grilli and Pratesi [49], Asparouhov [6], and Rabe-Hesketh and Skrondal [94] for a discussion of different definitions of weights and empirical studies of their properties. Chantala et al. [20] provide software that computes appropriate multilevel sampling weights for usage in several software packages.

Let us now assume that we have a set of weights, and we would like to compute the weighted version of the within-groups OLS estimate $\boldsymbol{b}_j$. The formula for the latter can be written as

$$\boldsymbol{b}_j \stackrel{\Delta}{=} (\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j'\boldsymbol{y}_j = \left(\frac{1}{n_j}\sum_{i=1}^{n_j}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}'\right)^{-1}\left(\frac{1}{n_j}\sum_{i=1}^{n_j}\boldsymbol{x}_{ij}y_{ij}\right).$$

Clearly, each of the two factors contains some kind of average, so that the analogy with average height mentioned above gives the following weighted estimate:

$$\begin{aligned}\boldsymbol{b}_{wj} &\stackrel{\Delta}{=} \left(\frac{\sum_{i=1}^{n_j} w_{i|j}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}'}{\sum_{i=1}^{n_j} w_{i|j}}\right)^{-1}\left(\frac{\sum_{i=1}^{n_j} w_{i|j}\boldsymbol{x}_{ij}y_{ij}}{\sum_{i=1}^{n_j} w_{i|j}}\right)\\ &= \left(\sum_{i=1}^{n_j} w_{i|j}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}'\right)^{-1}\left(\sum_{i=1}^{n_j} w_{i|j}\boldsymbol{x}_{ij}y_{ij}\right)\\ &= (\boldsymbol{X}_j'\boldsymbol{\mathcal{W}}_j\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j'\boldsymbol{\mathcal{W}}_j\boldsymbol{y}_j,\end{aligned}$$

where $\boldsymbol{\mathcal{W}}_j$ (not to be confused with $\boldsymbol{W}_j$) is the diagonal matrix with elements $w_{i|j}$ on its diagonal. A corresponding suitable estimator of σ_j^2 is obtained by a properly scaled version of the weighted sum of squared residuals. For the *unbiased* estimator, the denominator in this is a bit more complicated than in the unweighted case. The resulting formula is

$$s_{wj}^2 \stackrel{\Delta}{=} (\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{b}_{wj})'\boldsymbol{\mathcal{W}}_j(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{b}_{wj})/(n_j^* - p^*),$$

where

$$\begin{aligned}n_j^* &\stackrel{\Delta}{=} \sum_{i=1}^{n_j} w_{i|j} = \operatorname{tr}\boldsymbol{\mathcal{W}}_j\,,\\ p^* &\stackrel{\Delta}{=} \operatorname{tr}\big[(\boldsymbol{X}_j'\boldsymbol{\mathcal{W}}_j\boldsymbol{X}_j)^{-1}(\boldsymbol{X}_j'\boldsymbol{\mathcal{W}}_j^2\boldsymbol{X}_j)\big].\end{aligned}$$

Then, paraphrasing our earlier discussion and simplifying somewhat, for estimating $\boldsymbol{\gamma}$, least squares loss functions incorporating sampling weights can be defined as

$$\rho_w(\boldsymbol{\gamma}) \triangleq \sum_{j=1}^m w_j(\boldsymbol{b}_{wj} - \boldsymbol{Z}_j\boldsymbol{\gamma})'\boldsymbol{B}_j^{-1}(\boldsymbol{b}_{wj} - \boldsymbol{Z}_j\boldsymbol{\gamma}),$$

leading to the estimators

$$\hat{\underline{\gamma}}_{w,\boldsymbol{B}} \triangleq \left(\sum_{j=1}^m w_j \boldsymbol{Z}_j'\boldsymbol{B}_j^{-1}\boldsymbol{Z}_j\right)^{-1} \sum_{j=1}^m w_j \boldsymbol{Z}_j'\boldsymbol{B}_j^{-1}\underline{\boldsymbol{b}}_{wj}.$$

Because

$$\boldsymbol{W}_{wj} \triangleq \mathrm{Cov}(\underline{\boldsymbol{b}}_{wj}) = \boldsymbol{\Omega} + \sigma_j^2(\boldsymbol{X}_j'\boldsymbol{\mathcal{W}}_j\boldsymbol{X}_j)^{-1}(\boldsymbol{X}_j'\boldsymbol{\mathcal{W}}_j^2\boldsymbol{X}_j)(\boldsymbol{X}_j'\boldsymbol{\mathcal{W}}_j\boldsymbol{X}_j)^{-1},$$

the covariance matrices of these least squares estimators are

$$\left(\sum_{j=1}^m w_j \boldsymbol{Z}_j'\boldsymbol{B}_j^{-1}\boldsymbol{Z}_j\right)^{-1} \left(\sum_{j=1}^m w_j^2 \boldsymbol{Z}_j'\boldsymbol{B}_j^{-1}\boldsymbol{W}_{wj}\boldsymbol{B}_j^{-1}\boldsymbol{Z}_j\right) \left(\sum_{j=1}^m w_j \boldsymbol{Z}_j'\boldsymbol{B}_j^{-1}\boldsymbol{Z}_j\right)^{-1}.$$

The estimators corresponding to the 1-step and 2-step OLS estimators are obtained by choosing $\boldsymbol{B}_j = (\boldsymbol{X}_j'\boldsymbol{\mathcal{W}}_j\boldsymbol{X}_j)^{-1}$ and $\boldsymbol{B}_j = \boldsymbol{I}$, respectively. The most logical analog of the WLS estimator seems to be the one based on $\boldsymbol{B}_j = \boldsymbol{W}_{wj}$, but the optimality properties of the unweighted version do not hold and the covariance matrix does not simplify considerably. A different WLS estimator for data with sampling weights,

$$\begin{aligned}\hat{\underline{\gamma}}_{w,\mathrm{KR}} &\triangleq \left(\sum_{j=1}^m w_j \boldsymbol{U}_j'\boldsymbol{V}_j^{-1}\boldsymbol{U}_j\right)^{-1} \sum_{j=1}^m w_j \boldsymbol{U}_j'\boldsymbol{V}_j^{-1}\underline{\boldsymbol{y}}_j \\ &= \left(\sum_{j=1}^m w_j \boldsymbol{Z}_j'\boldsymbol{W}_j^{-1}\boldsymbol{Z}_j\right)^{-1} \sum_{j=1}^m w_j \boldsymbol{Z}_j'\boldsymbol{W}_j^{-1}\underline{\boldsymbol{b}}_j,\end{aligned}$$

using the unweighted within-groups estimates $\boldsymbol{b}_j$ and $\boldsymbol{W}_j$, was proposed by Kovačević and Rai [66]. This also does not have the optimality properties of the WLS estimator without sampling weights.

Generally, we need an estimate of $\boldsymbol{\Omega}$ as well. The estimators discussed earlier can be adapted relatively straightforwardly, but we omit this here, with the exception of a general treatment of ML with sampling weights.

The loglikelihood function for a two-level model that is not necessarily linear can be written as

$$\mathcal{L} = \sum_{j=1}^m \log \int \exp(\mathcal{L}_{j|\boldsymbol{\delta}_j})\, f_{\boldsymbol{\delta}}(\boldsymbol{\delta}_j)\, \mathrm{d}\boldsymbol{\delta}_j,$$

where we have suppressed the dependence on the parameter vector $\boldsymbol{\theta}$. The function $f_{\boldsymbol{\delta}}(\cdot)$ is the density function of $\underline{\boldsymbol{\delta}}_j$ and $\mathcal{L}_{j|\boldsymbol{\delta}_j}$ is the loglikelihood of the j-th group conditional on $\boldsymbol{\delta}_j$. Thus,

$$\mathcal{L}_{j|\boldsymbol{\delta}_j} = \sum_{i=1}^{n_j} \log f_{y\,|\,\boldsymbol{\delta}}(y_{ij} \mid \boldsymbol{\delta}_j)$$

in obvious notation. From this form, the adaptation for sampling weights is straightforward, leading to

$$\mathcal{L}_{w,j|\boldsymbol{\delta}_j} \stackrel{\Delta}{=} \sum_{i=1}^{n_j} w_{i|j} \log f_{y\,|\,\boldsymbol{\delta}}(y_{ij} \mid \boldsymbol{\delta}_j),$$
$$\mathcal{L}_w \stackrel{\Delta}{=} \sum_{j=1}^{m} w_j \log \int \exp(\mathcal{L}_{w,j|\boldsymbol{\delta}_j})\, f_{\boldsymbol{\delta}}(\boldsymbol{\delta}_j)\, \mathrm{d}\boldsymbol{\delta}_j = \sum_{j=1}^{m} w_j \mathcal{L}_{wj},$$

with $\mathcal{L}_{wj}$ implicitly defined. Thus, the first-order condition for the ML estimator with sampling weights is

$$\sum_{j=1}^{m} w_j \frac{\partial \mathcal{L}_{wj}}{\partial \boldsymbol{\theta}} = \boldsymbol{\emptyset}, \tag{1.32}$$

so that, adapting (1.30), the covariance estimate for the resulting estimator $\underline{\hat{\boldsymbol{\theta}}}$ becomes

$$\left(\sum_{j=1}^{m} w_j \frac{\partial^2 \mathcal{L}_{wj}}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}})\right)^{-1} \left(\sum_{j=1}^{m} w_j^2 \frac{\partial \mathcal{L}_{wj}}{\partial \boldsymbol{\theta}}(\hat{\boldsymbol{\theta}})\, \frac{\partial \mathcal{L}_{wj}}{\partial \boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}})\right) \left(\sum_{j=1}^{m} w_j \frac{\partial^2 \mathcal{L}_{wj}}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}})\right)^{-1}.$$

Unlike the covariance matrix without sampling weights, this formula does not simplify considerably even if all model assumptions are met. Thus, this illustrates that the resulting estimators are not proper ML estimators and the weighted loglikelihood function is not a proper loglikelihood. The estimators can, however, be viewed as generalized estimating equation (GEE) estimators based on the estimating equations (1.32) and, under weak regularity conditions, have desirable statistical properties (consistency, asymptotic normality). From this theory, it also follows that it is immaterial whether the weights are predetermined (by the sampling scheme) or estimated afterward (because of differential nonresponse), in which case they would be random variables. The estimating equations are still valid, unless the nonresponse is related to the dependent variable of interest ("nonignorable"), in which case analyzing the data becomes much more complicated and perhaps consistent estimators do not exist.

Of course, the formulas for the ML estimators with sampling weights simplify considerably for the linear multilevel model. This is straightforward and we do not give the expresssions here.

More extensive discussions of how to treat sampling weights in survey data in general and with multilevel models in particular can be found in Skinner [108], Pfeffermann [87], Pfeffermann et al. [88], and Asparouhov [5, 6].

1.9 A School Effects Example

In this section, we apply some of the techniques discussed in this chapter by analyzing the well-known NELS-88 data. These have been used to illustrate multilevel techniques by several authors and, of course, they have been used in substantive research as well.

The part of the NELS data that we use contains information about the score on a mathematics test, which will be our dependent variable, and the amount of time spent on homework, which will be our level-1 explanatory variable, and the student-teacher ratio of the school, which will be our level-2 explanatory variable. The math test score is a continuous variable having a sample average of 51, with a range of 27–71. Homework is coded from 0 = "None" to 7 = "10 or more hours per week". This is a slightly nonlinear transformation of the hours, reflecting expected diminishing returns from additional hours of homework. Both the average and the median of this variable are 2. The student-teacher ratio varies from 10 to 30, with mean and median approximately equal to 17. The data set consists of 21,580 students in 1003 schools, so the average number of observations per school is about 22. The number of observations per school varies from 1 to 67.

Kreft and de Leeuw [67] have previously analyzed this data set with multilevel analysis. We base our analyses on the model they describe in their Chapter 4. However, whereas their goal is to discuss different model specifications and the choice between them, we focus on comparing results for the same model obtained with different estimators.

In line with the description in this chapter, we start by computing the within-school regressions. This immediately illustrates a drawback of our focus on two-step estimators: In 10 schools, the within-groups regression coefficients $\boldsymbol{b}_j$ and/or the within groups residual variance s_j^2 cannot be computed because the sample size is too small ($n_j \leq p = 2$) or because $\boldsymbol{X}_j$ is not of full column rank, which is presumably also due to small sample size. Thus, we drop these 10 schools and proceed with the 993 remaining schools, leaving us with 21,558 observations. We do not expect that this seriously affects the results, and this is confirmed by the closeness of our results with the corresponding ones in Kreft and de Leeuw [67]. However, this also indicates that models that use different within-groups residual variances (σ_j^2) will not reliably estimate these parameters for schools with small numbers of observations.

After these disclaimers, we report the within-schools results for the first 30 successfully analyzed schools in Table 1.1. It shows considerable variation both in the regression coefficients and in the residual variances. This is corroborated by summary statistics for the whole sample: The within-groups intercept varies from 34 to 72, with mean and median approximately equal to 48, and the regression coefficient for homework varies from −12 to +15, with mean and median equal to 1.3, but more than 75% are positive. Finally, the

Table 1.1 Within-school statistics for the first 30 successfully analyzed schools: school identifier, number of pupils, student-teacher ratio, regression coefficients, and residual variance.

School ID	Observations	S-t ratio	Regression coefficient		Residual variance
			Constant	Homework	
1249	24	21	54.0969	−0.5760	66.6295
1755	14	16	45.9339	0.3330	60.6991
1806	15	25	45.8242	3.0579	70.4722
1846	36	28	45.3300	1.5674	62.4661
2114	19	13	57.5974	−0.6658	83.7773
2335	19	11	60.0461	0.5249	16.1703
2666	20	14	43.0026	3.1134	69.1364
2759	17	10	57.3730	−2.8981	86.0793
2861	21	17	52.5275	2.6298	73.8099
2888	20	30	53.5131	0.4496	71.1451
2988	23	22	51.0928	0.5839	99.0531
6043	10	23	57.0538	0.5509	54.7340
6044	24	23	55.4732	0.1090	65.2169
6053	44	18	51.6696	2.0880	75.1713
6091	8	22	47.7969	−0.3928	108.5720
6185	3	19	47.9300	0.7850	41.3438
6327	8	23	63.8000	−8.6350	25.8185
6358	10	28	60.6133	0.5409	16.9813
6375	4	20	57.5608	0.4358	21.8832
6420	7	25	53.0421	0.2061	70.3876
6442	11	12	48.8171	0.1168	101.7292
6467	5	19	41.0639	6.9128	11.8384
6518	21	29	60.2006	0.9153	64.9436
6631	5	20	68.5750	−7.4025	40.5725
6641	29	15	50.2446	1.5950	70.2012
6656	4	16	37.7940	3.9710	10.9923
6738	3	26	54.9100	−6.0000	10.7648
6868	18	13	52.3958	0.9523	63.7598
7000	24	13	41.6905	1.2020	72.8585
7011	20	24	45.9697	1.6501	62.8256

residual variance varies from 5 to 180, with mean and median approximately equal to 71. It is the goal of the second step of the analysis to model at least some of the variation in the regression coefficients.

Of course, a negative coefficient for time spent on homework does not make sense substantively. Rather, in addition to the possibility of sheer random fluctuation, this points to a possible endogeneity problem, caused by students who have more problems with mathematics spending more time on their homework. That is, it may be the result of a partial reversal of causality. For

the analysis here, we will ignore this possibility, given that we are primarily interested in differences between estimators.

We proceed by computing the one-step and two-step OLS estimates of the regression coefficients γ. These are reported in the first two columns of Table 1.2. The estimates are in the first panel, model-based standard errors (computed using the de Leeuw and Kreft [28] estimate of $\boldsymbol{\Omega}$) in the second panel, and robust standard errors in the third panel. Unlike a similar comparison for different data in de Leeuw and Kreft [28], we see some important differences between these estimates. The estimated main effect of the student-teacher ratio is twice as large for the two-step estimator, whereas the main effect of homework is less than half as large and the interaction term is also considerably less important, even statistically insignificant.

By using the within-groups and two-step OLS estimates, we can estimate $\boldsymbol{\Omega}$ by the method of de Leeuw and Kreft [28] discussed above. The estimate is denoted by DLK in Table 1.3. Fortunately, this is positive definite, so we do not encounter the problems faced by de Leeuw and Kreft for their example. Thus, we can use this estimate to compute the WLS estimates of γ. They are given in the third column of Table 1.2. They are very similar to the two-step estimates. As mentioned above, the estimate of $\boldsymbol{\Omega}$ is also be used in computing the model-based standard errors of the one-step and two-step OLS and WLS estimates, which are given in the second panel of Table 1.2. The third panel contains standard errors obtained from the cluster-robust covariance matrices.

Table 1.2 Estimates of fixed regression coefficients for the NELS-88 data and their standard errors.

	OLS (1-step)	OLS (2-step)	WLS (DLK)	FIML (1 σ)	REML (1 σ)	FIML (sep. σ's)	REML (sep. σ's)
				Estimates			
Constant	49.1477	52.1147	52.1062	51.4428	51.4434	51.9983	51.9988
S-t ratio	−0.1113	−0.2217	−0.2290	−0.2006	−0.2006	−0.2242	−0.2242
Homework	2.8520	1.2834	1.2785	1.5272	1.5272	1.3557	1.3561
hw × ratio	−0.0522	−0.0003	0.0058	−0.0030	−0.0030	0.0028	0.0028
				Model-based standard errors			
Constant	0.7857	0.7303	0.6913	0.7003	0.7011	0.7307	0.7314
S-t ratio	0.0428	0.0398	0.0378	0.0382	0.0382	0.0399	0.0400
Homework	0.2642	0.2362	0.1875	0.1823	0.1825	0.1781	0.1783
hw × ratio	0.0142	0.0127	0.0103	0.0100	0.0100	0.0098	0.0099
				Robust standard errors			
Constant	0.8176	0.8287	0.8862	0.8077	0.8049	0.8751	0.8639
S-t ratio	0.0433	0.0437	0.0469	0.0428	0.0427	0.0462	0.0457
Homework	0.2166	0.2225	0.1922	0.1828	0.1782	0.1961	0.1767
hw × ratio	0.0117	0.0118	0.0105	0.0098	0.0097	0.0105	0.0097

Next, we compute ML estimates. There are four of them: FIML and REML, each with a common variance parameter σ^2 or with separate variances σ_j^2. The results for the fixed coefficients are listed in the last four columns of Table 1.2. As argued before, these REML results are better called "WLS based on REML estimates of the variance parameters", but for convenience we call them REML here, and similarly WLS based on the DLK variance parameter estimates will be simply called WLS. The model-based standard errors for the ML estimators are obtained from the information matrix, whereas the robust standard errors are obtained from the cluster-robust covariance matrices described above. An exception is formed by the robust standard errors accompanying FIML with separate residual variances. These have been computed by formulas based on a combination of within-groups and between-groups asymptotics, as briefly mentioned but not worked out above (details are available upon request). This is intended to avoid the problems with the cluster-based estimator of the variance of the first derivatives of the loglikelihood, because its σ part is based on only 1 independent observation. However, the within-groups asymptotics involve sample fourth-order moments, which are highly inaccurate for the many small within-groups sample sizes. Nevertheless, the numerical results are similar to the ones for the other ML estimators, and also very similar to the two-step OLS and WLS results.

Note that the robust s.e.'s of the REML estimator are simply the WLS formulas, and thus are not affected by this problem. Given that the FIML estimators of $\boldsymbol{\gamma}$ are also WLS estimators, based on the FIML estimates of the variance parameters, we could have done the same for FIML. On the other hand, these WLS-based variance estimates essentially ignore any variability in the estimators of the variance parameters, which is also only asymptotically warranted.

The DLK and ML estimates of the elements of the level-2 covariance matrix $\boldsymbol{\Omega}$ are given in Table 1.3. The ML estimates using a single residual variance parameter are very similar to the DLK estimates (which are, incidentally, based on separate residual variances). The standard errors are a bit smaller, reflecting the higher precision of ML. When separate residual variances are estimated with ML, the estimates of $\boldsymbol{\Omega}$ are noticeably larger.

For both ML estimators with a single residual variance parameter, the estimate of σ^2 is 71.74 with a model-based standard error of 0.72 and a robust standard error of 0.85. The value of 71 corresponds closely with the average of the within-groups residual variance estimates.

For FIML with separate variances, the estimates of the residual variances vary from 8 to 161, with mean and median again approximately equal to 71. Similarly, for REML with separate variances, the estimates of the residual variances vary from 8 to 157, with mean and median also approximately equal to 71. This range is slightly narrower than the range of the within-groups

Table 1.3 Estimates of level-2 variance parameters for the NELS-88 data and their standard errors.

	DLK	FIML (1 σ)	REML (1 σ)	FIML (sep. σ's)	REML (sep. σ's)
			Estimates		
Constant, constant	23.9283	23.2633	23.3326	27.8982	27.9745
Homework, constant	−0.9319	−0.9105	−0.9197	−1.6088	−1.6197
Homework, homework	0.8678	0.5190	0.5243	0.6828	0.6878
			Model-based standard errors		
Constant, constant	1.8298	1.5125	1.5172	1.6826	1.6826
Homework, constant	0.6159	0.3138	0.3149	0.3296	0.3296
Homework, homework	0.3691	0.0993	0.0998	0.0971	0.0971
			Robust standard errors		
Constant, constant	—	1.5646	1.5591	1.7509	1.7509
Homework, constant	—	0.2983	0.2931	0.3197	0.3197
Homework, homework	—	0.1048	0.1047	0.1262	0.1262

Note: Robust standard errors are not available for the DLK [28] estimator.

estimates of the residual variances, but otherwise seems to confirm that the residual variances are not equal.

We can compute a likelihood ratio test statistic comparing the model with a common residual variance with the model with separate variances. For both FIML and REML, its value is approximately 1500, with 992 degrees of freedom, which gives a hugely significant p-value of approximately 2.2×10^{-23}. Even though the chi-square approximation is possibly inaccurate with such a large number of degrees of freedom and such small within-groups sample sizes, it clearly points in the direction of heterogeneous variances.

This leaves us with the conclusion that a model with a common variance is likely misspecified and a model with separate variances cannot be estimated reliably. Thus, this is a case in point for a more genuine multilevel approach in which the residual variance is modeled with a systematic part and a random residual, as suggested earlier.

Fortunately, however, the estimates and standard errors of the fixed coefficients, and to a lesser degree also the results for the level-2 covariance matrix, appear fairly insensitive to the specification of the level-1 random part. Thus, substantive conclusions would also be largely unaffected by this issue.

Clearly, this single empirical example is only an illustration and cannot be viewed as representative of all multilevel analyses. Many more examples, showing various issues in model specification and estimation, are discussed in detail in the textbooks [46, 59, 67, 76, 89, 101, 110, 111], the program manuals, and many empirical articles cited here and in the mentioned textbooks. Finally,

the remaining chapters of this Handbook contain many empirical applications as well, although for more complicated models.

1.10 Final Remarks

In this final section, we would like to briefly mention a few topics that have not been addressed in the previous sections. The first is *hypothesis tests*. Of course, this is one of the main topics of statistics (and typically the one that gives statistics its bad reputation among students in the social sciences). However, there is almost nothing that is specific to multilevel analysis. Thus, the general theory of hypothesis testing as presented in, e.g., Cameron and Trivedi [17, Chapter 7], and in particular, the well-known Wald, likelihood ratio, and Lagrange multiplier tests, can be directly applied. The only thing worth mentioning is that the REML loglikelihood cannot be used to test hypotheses concerning $\boldsymbol{\gamma}$, i.e., exclusion of certain variables from the fixed part of the model, because when viewed as a proper loglikelihood, it does not contain $\boldsymbol{\gamma}$.

More generally, model fit is an important subject. In addition to formal hypothesis tests, this typically involves certain more descriptive indexes of model fit, like R^2 in linear regression. Several such indexes have been proposed for multilevel analysis, but these tend to have serious drawbacks. Sometimes it is not guaranteed that the fit index improves as variables (or, more generally, parameters) are added to the model, whereas other fit indexes do not have a clear intuitive interpretation. Thus, the literature does not seem to have converged on this topic. See, e.g., Snijders and Bosker [111, Chapter 7], Hox [59, Section 4.4], Spiegelhalter et al. [112], Xu [128], and Gelman and Pardoe [41] for some proposed indexes and their properties. A systematic approach to diagnosing model (mis)specification, directed at various directions of misspecification, is given in Chapter 3 of this volume.

An important issue in multilevel model specification is *centering*. In social science data, variables typically do not have a natural zero point, and even if there is a natural zero, it may still not be an important baseline value. Therefore, in regression analysis and other multivariate statistical analysis methods, variables are often centered, so that the zero point is the sample average, which *is* an important baseline value. This tends to ease the interpretation of the parameters, especially the intercept, and it sometimes has some computational advantages as well. This practice has also been advocated for multilevel analysis, but the consequences for multilevel analysis are not as innocuous as for ordinary linear regression analysis. Moreover, in multilevel analysis, there are two possibilities for centering the data. The first is *grand mean centering*, i.e., the sample average of all observations is subtracted, and the second is *within-groups centering*, where the sample average of only the observations within the same group is subtracted. Generally, grand mean

centering does not change the model and is thus innocuous, but within-groups centering implicitly changes the model that is estimated, unless the sample averages of all level-1 predictor variables are included as level-2 predictors. For an extensive analysis, see Kreft et al. [68], Van Landeghem et al. [119], de Leeuw [27], and the references therein.

We close by noting that the quality of every data analysis crucially depends on the quality of the data. Most issues in data quality are not specific to multilevel analysis and are thus not discussed here. One important aspect, however, is the *sampling design*. Because a multilevel data set has observations at different levels, deciding on issues like sample size and randomization becomes more complicated than with single-level data. This subject is treated in detail in Chapter 4 in this volume.

Appendix

1.A Notational Conventions

This appendix describes the notation used in this chapter. The notation throughout this Handbook has been made as consistent as possible, so that this appendix also serves as a reference for the other chapters. However, the reader may occasionally discern slight differences in notation between the chapters.

1.A.1 Existing Notation

We used the most common books on mixed, random coefficient, and multilevel models to find a compromise notation [24, 46, 67, 76, 89, 101, 111]. There is a substantial agreement on notation in these books, although there are of course many differences of detail.

1.A.2 Matrices and Vectors

Matrices are boldface capitals; vectors are lowercase bold. In general, we use Greek symbols for unknowns and unobservables, such as parameters or latent variables (disturbances, variance components).

As another convention, we write $\boldsymbol{X}[n,r]$ for "$\boldsymbol{X}$ is an $n \times r$ matrix" and $\boldsymbol{y}[n]$ for "$\boldsymbol{y}$ is an n-element vector". Also, $\boldsymbol{X} = (x_{ij})$ is used to define a matrix in terms of its elements.

Two special matrix symbols we use are $\oplus$ for the *direct sum* and $\otimes$ for the *direct* (or *Kronecker*) *product*. If $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_p$ are matrices, with $\boldsymbol{A}_s[n_s, m_s]$, then the direct sum is the $\sum_{s=1}^{p} n_s \times \sum_{s=1}^{p} m_s$ matrix

$$\bigoplus_{s=1}^{p} \boldsymbol{A}_s = \boldsymbol{A}_1 \oplus \cdots \oplus \boldsymbol{A}_p = \begin{pmatrix} \boldsymbol{A}_1 & \emptyset & \emptyset & \dots & \emptyset \\ \emptyset & \boldsymbol{A}_2 & \emptyset & \cdots & \emptyset \\ \emptyset & \emptyset & \boldsymbol{A}_3 & \cdots & \emptyset \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \emptyset & \emptyset & \emptyset & \cdots & \boldsymbol{A}_p \end{pmatrix},$$

where $\emptyset$ denotes a (sub-)matrix with all elements equal to zero. The direct product is a $\prod_{s=1}^{p} n_s \times \prod_{s=1}^{p} m_s$ matrix, which we can best define recursively starting with two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. If $\boldsymbol{A}$ is $n \times m$, then

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{pmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & a_{13}\boldsymbol{B} & \cdots & a_{1m}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & a_{23}\boldsymbol{B} & \cdots & a_{2m}\boldsymbol{B} \\ a_{31}\boldsymbol{B} & a_{32}\boldsymbol{B} & a_{33}\boldsymbol{B} & \cdots & a_{3m}\boldsymbol{B} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1}\boldsymbol{B} & a_{n2}\boldsymbol{B} & a_{n3}\boldsymbol{B} & \cdots & a_{nm}\boldsymbol{B} \end{pmatrix}$$

and, by recursion,

$$\bigotimes_{s=1}^{p} \boldsymbol{A}_s = \boldsymbol{A}_1 \otimes (\boldsymbol{A}_2 \otimes \cdots \otimes \boldsymbol{A}_p).$$

Superscripted delta is the *Kronecker delta*, i.e.,

$$\delta^{st} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases}$$

The identity matrix is $\boldsymbol{I}$, a vector with all elements equal to 1 is $\mathbf{1}$. The matrix $\boldsymbol{E}$ has all elements equal to 1. The size of these matrices and vectors will often be clear from the context. If we need to be explicit, we can always write, for instance, $\boldsymbol{E}[n, m]$, but we also use the forms $\boldsymbol{I}_n$ and $\mathbf{1}_n$. Unit vectors $\boldsymbol{e}_i$ have all elements equal to zero, except for element i, which is equal to 1. Thus, $\mathbf{1}$ is the sum of the $\boldsymbol{e}_i$.

1.A.3 Special Symbols

We use the following special symbols:

$\stackrel{\Delta}{=}$	is defined as
$\sim$	is distributed as
$\mathcal{N}$	normal distribution
$\stackrel{\mathcal{L}}{\Longrightarrow}$	convergence in law (distribution)
$\stackrel{a.d.}{=}$	has the same asymptotic distribution
$\stackrel{\mathcal{P}}{\Longrightarrow}$	convergence in probability
$\stackrel{iid}{\sim}$	i.i.d. with given distribution

1.A.4 Underlining Random Variables

A non-standard part of our notation is that we *underline random variables* [28]. Thus, vector or matrix random variables are both underlined and bold.

The advantage of distinguishing between random variables and fixed known or unknown constants in the context of mixed models is clear. We use constants (the design matrix, unknown parameters) and random variables (the outcome variables, of which we observe a realization, and the random effects, which we do not observe at all). We also estimate parameters. Estimates are fixed values, realization of estimators, which are random variables. Underlining gives us an extra alphabet, it also gives us a method to indicate how constants and random variables are related, because we can use y for a realization of $\underline{y}$. The advantages of underlining, known as the *Dutch Convention* or *Van Dantzig Convention*, are discussed in more detail in Hemelrijk [58].

As a simple example, the classical linear model is

$$\underline{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\beta} + \underline{\boldsymbol{\epsilon}},$$

with

$$\underline{\boldsymbol{\epsilon}} \sim \mathcal{N}(\boldsymbol{\emptyset}, \sigma^2\boldsymbol{I}).$$

Thus,

$$\underline{\boldsymbol{y}} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}).$$

We observe $\boldsymbol{y}$ and $\boldsymbol{X}$, and we compute

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}, \tag{1.33}$$

which is a realization of a random variable $\underline{\hat{\boldsymbol{\beta}}}$, satisfying

$$\underline{\hat{\boldsymbol{\beta}}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}).$$

It obviously makes sense to write $E(\underline{\hat{\boldsymbol{\beta}}}) = \boldsymbol{\beta}$, and it does not make sense to write $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.

Equation (1.33) also illustrates the convention of writing the estimate of a parameter by putting a hat on the parameter symbol. We also use this convention for "estimating" a random component, for instance,

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}.$$

For conditional expectations, we can both have $E(\underline{x} \mid y)$ and $E(\underline{x} \mid \underline{y})$, because we can condition on both a random variable and its realization. The first expression defines a deterministic function of y, the second a function of $\underline{y}$, i.e., a random variable.

It is important to emphasize some basic consequences of our conventions. Anything we actually compute cannot be underlined, because we only compute with realizations, not with random variables. Anything that is underlined is by definition part of a statistical model, because it implies a framework of replication or a degree of belief. In Bayesian models, there will be more underlining than in empirical Bayes models, and empirical Bayes models have more underlining than classical frequentist models. Ultimately, of course, even fully Bayesian models will have fixed hyperparameters, because otherwise the specification of the model will never stop.

1.B Generic Numerical Optimization

The most common starting point for numerical optimization of a generic well-behaved function is a second-order Taylor series expansion around a point $\boldsymbol{\theta}_1$:

$$f(\boldsymbol{\theta}) = f_1 + \boldsymbol{g}_1'(\boldsymbol{\theta} - \boldsymbol{\theta}_1) + \tfrac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_1)'\boldsymbol{H}_1(\boldsymbol{\theta} - \boldsymbol{\theta}_1) + o\|\boldsymbol{\theta} - \boldsymbol{\theta}_1\|^2,$$

where f_1, $\boldsymbol{g}_1$, and $\boldsymbol{H}_1$ are the function $f(\cdot)$, its gradient $\boldsymbol{g}(\cdot)$ (vector of first partial derivatives with respect to $\boldsymbol{\theta}$), and its Hessian $\boldsymbol{H}(\cdot)$ (matrix of second partial derivatives with respect to $\boldsymbol{\theta}$), respectively, all evaluated in $\boldsymbol{\theta}_1$.

Thus, if we ignore the approximation error reflected by the last term, we find that the function is minimized for

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_1 - \boldsymbol{H}_1^{-1}\boldsymbol{g}_1,$$

provided that $\boldsymbol{H}_1$ is positive definite. Of course, in practice the approximation error is not zero, so that this does not minimize the loss function immediately. But we can assert that we have come closer and repeat the process, leading to the algorithm

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \boldsymbol{H}_i^{-1}\boldsymbol{g}_i,$$

where i denotes the iteration number. This algorithm defines the well-known *Newton-Raphson* method, also known simply as Newton's method. In practice, two modifications are often necessary to ensure that this algorithm works well. The first is that the *search direction* $-\boldsymbol{H}_i^{-1}\boldsymbol{g}_i$ is only guaranteed to point in the direction of smaller function values if $\boldsymbol{H}_i$ is positive definite. Hence, if the loss function is not globally convex, $\boldsymbol{H}_i$ may have to be modified in some iterations to ensure that it is positive definite. This is typically done by adding a positive multiple of the identity matrix until all eigenvalues are positive. The second modification that is often used is to insert a *step size* α_i, with which the search direction is multiplied, so that the algorithm becomes

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \alpha_i\boldsymbol{H}_i^{-1}\boldsymbol{g}_i, \tag{1.34}$$

where it is understood that $\boldsymbol{H}_i$ may be the modified version to make it positive definite. Even though it is guaranteed that the search direction points toward smaller function values, the unmodified update may "overshoot" if the function decreases slowly in the neighborhood of the current point, but then increases sharply. Therefore, the factor α_i is chosen such that the function value in the next point is smaller than in the current point. A value of α_i that ensures this always exists if $\boldsymbol{H}_i$ is positive definite and $\boldsymbol{g}_i$ is nonzero. Typically, one would start with $\alpha_i = 1$, halving step size until such a point is reached. The (modified) Newton-Raphson method is implemented in most general-purpose optimization functions.

There exist many alternative generic numerical optimization methods, most of which use the same form (1.34) of an iteration, but with $\boldsymbol{H}_i^{-1}$ replaced by another positive (semi)definite matrix. The reason for this is that it is often computationally demanding to compute $\boldsymbol{H}_i^{-1}$, and places a larger burden on the researcher and/or programmer, because the second derivatives have to be computed and programmed. In principle, these methods converge more slowly, because in the neighborhood of the minimum, the loss function is closely approximated by a quadratic function, so that Newton-Raphson converges very fast. In contrast, the *steepest descent* method, which simply replaces $\boldsymbol{H}_i^{-1}$ by the identity matrix, tends to converge extremely slowly. In many cases, however, the better alternative methods are not noticeably worse (in terms of speed and accuracy) than Newton-Raphson. A good and popular method is the BFGS method, which replaces $\boldsymbol{H}_i^{-1}$ by the matrix $\boldsymbol{G}_i$. The latter matrix is computed using the update formula

$$\boldsymbol{G}_{i+1} = (\boldsymbol{I} - \rho_i \triangle\boldsymbol{\theta}_i \triangle\boldsymbol{g}_i')\boldsymbol{G}_i(\boldsymbol{I} - \rho_i \triangle\boldsymbol{g}_i \triangle\boldsymbol{\theta}_i') + \rho_i \triangle\boldsymbol{\theta}_i \triangle\boldsymbol{\theta}_i',$$

where $\triangle\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}_i$, $\triangle\boldsymbol{g}_i = \boldsymbol{g}_{i+1} - \boldsymbol{g}_i$, and $\rho_i = 1/\triangle\boldsymbol{g}_i'\triangle\boldsymbol{\theta}_i$. Clearly, if $\boldsymbol{G}_i$ is positive semidefinite, then $\boldsymbol{G}_{i+1}$ is also positive semidefinite. Moreover, it can be proved that if $\boldsymbol{G}_i$ is positive definite, then $\boldsymbol{G}_{i+1}$ is also positive definite. Typically, the starting value $\boldsymbol{G}_0$ is the identity matrix, which is clearly positive definite, or an informed guess of $\boldsymbol{H}^{-1}$. When BFGS is applied to a (convex) quadratic function of an n-element vector $\boldsymbol{\theta}$, and the step size is chosen to minimize the function along the line defined by the update formula, the global minimum is attained in n iterations and $\boldsymbol{G}_{n+1} = \boldsymbol{H}^{-1}$ (which is a constant matrix). Therefore, unless the number of parameters is large, BFGS tends to converge quickly in the neighborhood of the minimum, where the loss function is approximately quadratic. The BFGS method is also implemented in most general-purpose optimization functions.

An extensive treatment of many generic numerical optimization procedures, including Newton-Raphson and BFGS, with derivations of their properties, can be found in Nocedal and Wright [86].

1.C Some Matrix Expressions

Here we collect some convenient results to deal with two-level linear models. The first two results have been known for a long time [26, 36, 117]. Proofs of the first three results are given, for example, in de Leeuw and Liu [31]. Many additional useful matrix results are provided by Wansbeek and Meijer [123, appendix A] and Harville [53].

Theorem 1.1 If $\boldsymbol{A} = \boldsymbol{B} + \boldsymbol{T}\boldsymbol{C}\boldsymbol{T}'$ with $\boldsymbol{A}$ and $\boldsymbol{B}$ positive definite, then

$$\log|\boldsymbol{A}| = \log|\boldsymbol{B}| + \log|\boldsymbol{C}| + \log|\boldsymbol{C}^{-1} + \boldsymbol{T}'\boldsymbol{B}^{-1}\boldsymbol{T}|.$$

If, in addition, $\boldsymbol{T}$ is of full column rank, then

$$\log|\boldsymbol{A}| = \log|\boldsymbol{B}| + \log|\boldsymbol{T}'\boldsymbol{B}^{-1}\boldsymbol{T}| + \log|\boldsymbol{C} + (\boldsymbol{T}'\boldsymbol{B}^{-1}\boldsymbol{T})^{-1}|.$$

Theorem 1.2 If $\boldsymbol{A} = \boldsymbol{B} + \boldsymbol{T}\boldsymbol{C}\boldsymbol{T}'$ with $\boldsymbol{A}$ and $\boldsymbol{B}$ positive definite, then

$$\boldsymbol{A}^{-1} = \boldsymbol{B}^{-1} - \boldsymbol{B}^{-1}\boldsymbol{T}(\boldsymbol{C}^{-1} + \boldsymbol{T}'\boldsymbol{C}^{-1}\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{B}^{-1}.$$

If, in addition, $\boldsymbol{T}$ is of full column rank, then

$$\begin{aligned}\boldsymbol{A}^{-1} = \boldsymbol{T}(\boldsymbol{T}'\boldsymbol{T})^{-1}(\boldsymbol{C} + (\boldsymbol{T}'\boldsymbol{B}^{-1}\boldsymbol{T})^{-1})^{-1}(\boldsymbol{T}'\boldsymbol{T})^{-1}\boldsymbol{T}' \\ + \{\boldsymbol{B}^{-1} - \boldsymbol{B}^{-1}\boldsymbol{T}(\boldsymbol{T}'\boldsymbol{B}^{-1}\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{B}^{-1}\}.\end{aligned}$$

Theorem 1.3 If $\boldsymbol{A} = \boldsymbol{B} + \boldsymbol{T}\boldsymbol{C}\boldsymbol{T}'$ with $\boldsymbol{A}$ and $\boldsymbol{B}$ positive definite,then

$$\boldsymbol{y}'\boldsymbol{A}^{-1}\boldsymbol{y} = \min_{\boldsymbol{x}}\{(\boldsymbol{y} - \boldsymbol{T}\boldsymbol{x})'\boldsymbol{B}^{-1}(\boldsymbol{y} - \boldsymbol{T}\boldsymbol{x}) + \boldsymbol{x}'\boldsymbol{C}^{-1}\boldsymbol{x}\}.$$

The fourth result was proved by de Hoog et al. [26] by letting $\boldsymbol{C}^{-1} \to \emptyset$ on both sides of Theorem 1.2.

Theorem 1.4 If $\boldsymbol{B}$ is positive definite and $\boldsymbol{T}$ is of full column-rank, then

$$\boldsymbol{B}^{-1} - \boldsymbol{B}^{-1}\boldsymbol{T}(\boldsymbol{T}'\boldsymbol{B}^{-1}\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{B}^{-1} = (\boldsymbol{Q}\boldsymbol{B}\boldsymbol{Q})^{+},$$

where $\boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{T}(\boldsymbol{T}'\boldsymbol{T})^{-1}\boldsymbol{T}'$ and superscript $+$ denotes the Moore-Penrose inverse.

1.D The EM Algorithm

The EM algorithm of Dempster et al. [33] is a general method to optimize functions of the form $f(\boldsymbol{\theta}) = \log\int g(\boldsymbol{\theta}, \boldsymbol{z})\,\mathrm{d}\boldsymbol{z}$ over $\boldsymbol{\theta}$, where $g(\boldsymbol{\theta}, \boldsymbol{z}) > 0$ for all $\boldsymbol{\theta}$ and $\boldsymbol{z}$ in the domain. It is usually presented in probabilistic terminology, but the reason why it works is the concavity of the logarithm, which is obviously not a probabilistic result.

Define $h(\boldsymbol{\theta}) \stackrel{\Delta}{=} \int g(\boldsymbol{\theta}, \boldsymbol{z})\ \mathrm{d}\boldsymbol{z}$ and $k(\boldsymbol{z} \mid \boldsymbol{\theta}) \stackrel{\Delta}{=} g(\boldsymbol{\theta}, \boldsymbol{z})/h(\boldsymbol{\theta})$. Then, by the concavity of the logarithm, it follows from Jensen's inequality [96, p. 58] that for all $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$,

$$f(\boldsymbol{\theta}) \geq f(\tilde{\boldsymbol{\theta}}) + \int \log g(\boldsymbol{\theta}, \boldsymbol{z})\, k(\boldsymbol{z} \mid \tilde{\boldsymbol{\theta}})\ \mathrm{d}\boldsymbol{z} - \int \log g(\tilde{\boldsymbol{\theta}}, \boldsymbol{z})\, k(\boldsymbol{z} \mid \tilde{\boldsymbol{\theta}})\ \mathrm{d}\boldsymbol{z}, \quad (1.35)$$

with equality if and only if $g(\boldsymbol{\theta}, \boldsymbol{z}) = g(\tilde{\boldsymbol{\theta}}, \boldsymbol{z})$ almost everywhere.

In each iteration of the EM algorithm we take $\tilde{\boldsymbol{\theta}}$ to be our current best approximation to the optimum and improve it by maximizing the right-hand side of (1.35) over $\boldsymbol{\theta}$ for this given $\tilde{\boldsymbol{\theta}}$. In other words, we find $\boldsymbol{\theta}^{(i+1)}$ by maximizing

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(i)}) \stackrel{\Delta}{=} \int \log g(\boldsymbol{\theta}, \boldsymbol{z})\, k(\boldsymbol{z} \mid \boldsymbol{\theta}^{(i)})\ \mathrm{d}\boldsymbol{z}$$

over $\boldsymbol{\theta}$. The algorithm is monotone, in the sense that $f(\boldsymbol{\theta}^{(i+1)}) > f(\boldsymbol{\theta}^{(i)})$ and in many cases this is enough to guarantee (linear) convergence to a local maximum of $f(\cdot)$.

In the probabilistic interpretation, $f(\boldsymbol{\theta})$ is a loglikelihood function and EM stands for expectation-maximization. The E-step computes $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(i)})$, which is the conditional expectation of the *complete-data loglikelihood* $g(\boldsymbol{\theta}, \boldsymbol{z})$, given the observed data and the current parameter value $\boldsymbol{\theta}^{(i)}$, and the M-step maximizes the resulting function.

We can now apply the EM algorithm to the multilevel FIML loglikelihood. Here, $\underline{\boldsymbol{z}}$ consists of all the random effects $\underline{\boldsymbol{\delta}}_j$, and $\boldsymbol{\theta}$ is the usual parameter vector. The complete-data loglikelihood has the form

$$g(\boldsymbol{\theta}, \boldsymbol{\delta}) = \prod_{j=1}^{m} g_j(\boldsymbol{\theta}, \boldsymbol{\delta}_j),$$

where $g_j(\boldsymbol{\theta}, \boldsymbol{\delta}_j)$ is the joint density of $\underline{\boldsymbol{y}}_j$ and $\underline{\boldsymbol{\delta}}_j$. Using standard probability theory, we can write

$$\begin{aligned}
g_j(\boldsymbol{\theta}, \boldsymbol{\delta}_j) &= f_{\boldsymbol{\delta} \mid \boldsymbol{y}}(\boldsymbol{\delta}_j \mid \boldsymbol{y}_j) f_{\boldsymbol{y}}(\boldsymbol{y}_j), \\
h_j(\boldsymbol{\theta}) &\stackrel{\Delta}{=} \int g_j(\boldsymbol{\theta}, \boldsymbol{\delta}_j)\ \mathrm{d}\boldsymbol{\delta}_j = f_{\boldsymbol{y}}(\boldsymbol{y}_j), \\
k_j(\boldsymbol{\delta}_j \mid \boldsymbol{\theta}) &\stackrel{\Delta}{=} g_j(\boldsymbol{\theta}, \boldsymbol{\delta}_j)/h_j(\boldsymbol{\theta}) = f_{\boldsymbol{\delta} \mid \boldsymbol{y}}(\boldsymbol{\delta}_j \mid \boldsymbol{y}_j), \\
Q_j(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(i)}) &\stackrel{\Delta}{=} \int \log g_j(\boldsymbol{\theta}, \boldsymbol{\delta}_j)\, k_j(\boldsymbol{\delta}_j \mid \boldsymbol{\theta}^{(i)})\ \mathrm{d}\boldsymbol{\delta}_j\,, \\
Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(i)}) &= \sum_{j=1}^{m} Q_j(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(i)}).
\end{aligned}$$

The joint distribution of $\underline{\boldsymbol{y}}_j$ and $\underline{\boldsymbol{\delta}}_j$ is normal:

$$\begin{pmatrix} \underline{\boldsymbol{y}}_j \\ \underline{\boldsymbol{\delta}}_j \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{U}_j\boldsymbol{\gamma} \\ \boldsymbol{\emptyset} \end{pmatrix}, \begin{pmatrix} \boldsymbol{V}_j & \boldsymbol{X}_j\boldsymbol{\Omega} \\ \boldsymbol{\Omega}\boldsymbol{X}_j' & \boldsymbol{\Omega} \end{pmatrix} \right),$$

from which we obtain the conditional distribution of $\underline{\boldsymbol{\delta}}_j$ given $\boldsymbol{y}_j$ as

$$\underline{\boldsymbol{\delta}}_j \mid \boldsymbol{y}_j \sim \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j),$$

with

$$\begin{aligned} \boldsymbol{\mu}_j &= \boldsymbol{\Omega}\boldsymbol{X}_j'\boldsymbol{V}_j^{-1}(\boldsymbol{y}_j - \boldsymbol{U}_j\boldsymbol{\gamma}) = \boldsymbol{\Omega}\boldsymbol{W}_j^{-1}(\boldsymbol{b}_j - \boldsymbol{Z}_j\boldsymbol{\gamma}), \\ \boldsymbol{\Sigma}_j &= \boldsymbol{\Omega} - \boldsymbol{\Omega}\boldsymbol{X}_j'\boldsymbol{V}_j^{-1}\boldsymbol{X}_j\boldsymbol{\Omega} = \sigma_j^2\boldsymbol{\Omega}\boldsymbol{W}_j^{-1}(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}. \end{aligned}$$

By writing $g_j(\boldsymbol{\theta}, \boldsymbol{\delta}_j) = f_{\boldsymbol{y}\,|\,\boldsymbol{\delta}}(\boldsymbol{y}_j \mid \boldsymbol{\delta}_j) f_{\boldsymbol{\delta}}(\boldsymbol{\delta}_j)$, and observing that the marginal distribution of $\underline{\boldsymbol{\delta}}_j$ is normal with mean zero and covariance matrix $\boldsymbol{\Omega}$, and the conditional distribution of $\underline{\boldsymbol{y}}_j$ given $\boldsymbol{\delta}_j$ is normal with mean $\boldsymbol{U}_j\boldsymbol{\gamma} + \boldsymbol{X}_j\boldsymbol{\delta}_j$ and covariance matrix $\sigma_j^2\boldsymbol{I}_{n_j}$, we obtain, after some simplification,

$$\begin{aligned} \log g_j(\boldsymbol{\theta}, \boldsymbol{\delta}_j) = &-\frac{n_j + p}{2}\log(2\pi) - \frac{n_j}{2}\log\sigma_j^2 - \frac{1}{2\sigma_j^2}(n_j - p)s_j^2 \\ &- \frac{1}{2\sigma_j^2}(\boldsymbol{b}_j - \boldsymbol{Z}_j\boldsymbol{\gamma})'\boldsymbol{X}_j'\boldsymbol{X}_j(\boldsymbol{b}_j - \boldsymbol{Z}_j\boldsymbol{\gamma}) + \frac{1}{\sigma_j^2}(\boldsymbol{b}_j - \boldsymbol{Z}_j\boldsymbol{\gamma})'\boldsymbol{X}_j'\boldsymbol{X}_j\boldsymbol{\delta}_j \\ &- \tfrac{1}{2}\log|\boldsymbol{\Omega}| - \tfrac{1}{2}\operatorname{tr}\left[(\sigma_j^{-2}\boldsymbol{X}_j'\boldsymbol{X}_j + \boldsymbol{\Omega}^{-1})\boldsymbol{\delta}_j\boldsymbol{\delta}_j'\right]. \end{aligned}$$

The function $Q_j(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(i)})$ is obtained by integrating the product of this with $k_j(\boldsymbol{\delta}_j \mid \boldsymbol{\theta}^{(i)})$. That is, it is obtained as the expectation of $\log g_j(\boldsymbol{\theta}, \underline{\boldsymbol{\delta}}_j)$ when viewed as a function of the random variable $\underline{\boldsymbol{\delta}}_j$ that is normally distributed with mean $\boldsymbol{\mu}_j^{(i)}$ and covariance matrix $\boldsymbol{\Sigma}_j^{(i)}$, which are $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ evaluated in $\boldsymbol{\theta}^{(i)}$. For this distribution, we evidently have $E(\underline{\boldsymbol{\delta}}_j) = \boldsymbol{\mu}_j^{(i)}$ and $E(\underline{\boldsymbol{\delta}}_j\underline{\boldsymbol{\delta}}_j') = \boldsymbol{\Sigma}_j^{(i)} + \boldsymbol{\mu}_j^{(i)}\boldsymbol{\mu}_j^{(i)\prime}$, so that, after some simplification, we obtain

$$\begin{aligned} Q_j(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(i)}) = &\left(-\frac{n_j + p}{2}\log(2\pi)\right) - \tfrac{1}{2}\left(\log|\boldsymbol{\Omega}| + \operatorname{tr}\left[\boldsymbol{\Omega}^{-1}(\boldsymbol{\Sigma}_j^{(i)} + \boldsymbol{\mu}_j^{(i)}\boldsymbol{\mu}_j^{(i)\prime})\right]\right) \\ &- \frac{n_j}{2}\log\sigma_j^2 - \frac{1}{2\sigma_j^2}(n_j - p)s_j^2 - \frac{1}{2\sigma_j^2}\operatorname{tr}(\boldsymbol{X}_j'\boldsymbol{X}_j\boldsymbol{\Sigma}_j^{(i)}) \\ &- \frac{1}{2\sigma_j^2}(\boldsymbol{b}_j - \boldsymbol{\mu}_j^{(i)} - \boldsymbol{Z}_j\boldsymbol{\gamma})'\boldsymbol{X}_j'\boldsymbol{X}_j(\boldsymbol{b}_j - \boldsymbol{\mu}_j^{(i)} - \boldsymbol{Z}_j\boldsymbol{\gamma}). \end{aligned}$$

Consequently, the parameter values that optimize $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(i)})$ are

$$\begin{aligned} \boldsymbol{\Omega}^{(i+1)} &= \frac{1}{m}\sum_{j=1}^{m}(\boldsymbol{\Sigma}_j^{(i)} + \boldsymbol{\mu}_j^{(i)}\boldsymbol{\mu}_j^{(i)\prime}), \\ \boldsymbol{\gamma}^{(i+1)} &= \left(\sum_{j=1}^{m}\boldsymbol{Z}_j'\boldsymbol{X}_j'\boldsymbol{X}_j\boldsymbol{Z}_j\right)^{-1}\sum_{j=1}^{m}\boldsymbol{Z}_j'\boldsymbol{X}_j'\boldsymbol{X}_j(\boldsymbol{b}_j - \boldsymbol{\mu}_j^{(i)}), \\ (\sigma_j^2)^{(i+1)} &= \frac{1}{n_j}\left[(n_j - p)s_j^2 + \operatorname{tr}(\boldsymbol{X}_j'\boldsymbol{X}_j\boldsymbol{\Lambda}_j^{(i)})\right], \end{aligned}$$

or, instead of the latter,

$$(\sigma^2)^{(i+1)} = \frac{1}{n}\sum_{j=1}^{m}\left[(n_j - p)s_j^2 + \text{tr}(\boldsymbol{X}_j'\boldsymbol{X}_j\boldsymbol{\Lambda}_j^{(i)})\right],$$

where

$$\boldsymbol{\Lambda}_j^{(i)} \triangleq \boldsymbol{\Sigma}_j^{(i)} + (\boldsymbol{b}_j - \boldsymbol{\mu}_j^{(i)} - \boldsymbol{Z}_j\boldsymbol{\gamma}^{(i+1)})(\boldsymbol{b}_j - \boldsymbol{\mu}_j^{(i)} - \boldsymbol{Z}_j\boldsymbol{\gamma}^{(i+1)})'.$$

Note that when $\boldsymbol{\Omega}$ is not completely free (apart from the requirements of symmetry and positive definiteness, of course), then the M-step with respect to the parameters $\{\xi_g\}$ is nontrivial. We then need to minimize the function

$$F(\boldsymbol{\xi}) \triangleq \log|\boldsymbol{\Omega}| + \text{tr}\big(\boldsymbol{\Omega}^{-1}\boldsymbol{S}^{(i)}\big)$$

with respect to $\boldsymbol{\xi} \triangleq (\xi_1, \ldots, \xi_G)'$, where

$$\boldsymbol{S}^{(i)} \triangleq \frac{1}{m}\sum_{j=1}^{m}(\boldsymbol{\Sigma}_j^{(i)} + \boldsymbol{\mu}_j^{(i)}\boldsymbol{\mu}_j^{(i)\prime}).$$

Assuming (1.3), the first-order conditions are

$$\text{tr}[\boldsymbol{\Omega}^{-1}(\boldsymbol{S}^{(i)} - \boldsymbol{\Omega})\boldsymbol{\Omega}^{-1}\boldsymbol{C}_g] = 0.$$

Letting $\boldsymbol{C}^*$ be the matrix with g-th column equal to $\text{vec}(\boldsymbol{C}_g)$, these can be jointly written as

$$\boldsymbol{C}^{*\prime}(\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}^{-1})(\text{vec}\,\boldsymbol{S}^{(i)} - \boldsymbol{C}^*\boldsymbol{\xi}) = \emptyset,$$

which is a nonlinear equation that does not generally have a closed-form solution. However, it strongly suggests that one or more IGLS iterations of the form

$$\boldsymbol{\xi}^{(i+1,k+1)} = \big[\boldsymbol{C}^{*\prime}(\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}^{-1})\boldsymbol{C}^*\big]^{-1}\boldsymbol{C}^{*\prime}(\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}^{-1})\,\text{vec}\,\boldsymbol{S}^{(i)},$$

where in the right-hand side $\text{vec}\,\boldsymbol{S}^{(i)}$ is held fixed throughout these subiterations, but $\boldsymbol{\Omega}$ is the value from the previous (k-th) subiteration, should also increase the loglikelihood, so that full optimization in this step is not necessary.

References

1. L. S. Aiken and S. G. West. *Multiple Regression: Testing and Interpreting Interaction.* Sage Publications, Newbury Park, CA, 1991.

2. M. Aitkin and N. Longford. Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society, Series A*, 149:1–43, 1986. (with discussion)
3. L. Anselin. Spatial econometrics. In B. H. Baltagi, editor, *A Companion to Theoretical Econometrics*, pages 310–330. Blackwell, Malden, MA, 2001.
4. M. Arellano. *Panel Data Econometrics*. Oxford University Press, Oxford, UK, 2003.
5. T. Asparouhov. Sampling weights in latent variable modeling. *Structural Equation Modeling*, 12:411–434, 2005.
6. T. Asparouhov. General multi-level modeling with sampling weights. *Communications in Statistics—Theory & Methods*, 35:439–460, 2006.
7. D. Bates and D. Sarkar. *The lme4 Package*, 2006. URL `http://cran.r-project.org`
8. P. M. Bentler. *EQS 6 Structural Equations Program Manual*. Multivariate Software, Encino, CA, 2006.
9. H. M. Blalock. Contextual effects models: Theoretical and methodological issues. *Annual Review of Sociology*, 10:353–372, 1984.
10. R. J. Bosker, T. A. B. Snijders, and H. Guldemond. *PINT: Estimating Standard Errors of Regression Coefficients in Hierarchical Linear Models for Power Calculations. User's Manual Version 1.6*. University of Twente, Enschede, The Netherlands, 1999.
11. L. H. Boyd and G. R. Iversen. *Contextual Analysis: Concepts and Statistical Techniques*. Wadsworth, Belmont, CA, 1979.
12. M. W. Browne. Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 37:62–83, 1984.
13. A. S. Bryk and S. W. Raudenbush. *Hierachical Linear Models: Applications and Data Analysis Methods*. Sage, Newbury Park, CA, 1992.
14. L. Burstein. The analysis of multilevel data in educational research and evaluation. *Review of Research in Education*, 8:158–233, 1980.
15. L. Burstein, R. L. Linn, and F. J. Capell. Analyzing multilevel data in the presence of heterogeneous within-class regressions. *Journal of Educational Statistics*, 3:347–383, 1978.
16. F. M. T. A. Busing, E. Meijer, and R. Van der Leeden. *MLA: Software for MultiLevel Analysis of Data with Two Levels. User's Guide for Version 4.1*. Leiden University, Department of Psychology, Leiden, 2005.
17. A. C. Cameron and P. K. Trivedi. *Microeconometrics: Methods and Applications*. Cambridge University Press, Cambridge, UK, 2005.
18. G. Chamberlain. Panel data. In Z. Griliches and M. D. Intriligator, editors, *Handbook of Econometrics*, volume 2, pages 1247–1318. North-Holland, Amsterdam, 1984.
19. G. Chamberlain and E. E. Leamer. Matrix weighted averages and posterior bounds. *Journal of the Royal Statistical Society, Series B*, 38:73–84, 1976.
20. K. Chantala, D. Blanchette, and C. M. Suchindran. Software to compute sampling weights for multilevel analysis, 2006. URL `http://www.cpc.unc.edu/restools/data_analysis/ml_sampling_weights`

21. Y. F. Cheong, R. P. Fotiu, and S. W. Raudenbush. Efficiency and robustness of alternative estimators for two- and three-level models: The case of NAEP. *Journal of Educational and Behavioral Statistics*, 26:411–429, 2001.
22. J. S. Coleman, E. Q. Campbell, C. J. Hobson, J. McPartland, A. M. Mood, F. D. Weinfeld, and R. L. York. *Equality of Educational Opportunity*. U.S. Government Printing Office, Washington, DC, 1966.
23. D. R. Cox. Interaction. *International Statistical Review*, 52:1–31, 1984.
24. M. Davidian and D. M. Giltinan. *Nonlinear Models for Repeated Mesurement Data*. Chapman & Hall, London, 1995.
25. R. Davidson and J. G. MacKinnon. *Estimation and Inference in Econometrics*. Oxford University Press, Oxford, UK, 1993.
26. F. R. de Hoog, T. P. Speed, and E. R. Williams. On a matrix identity associated with generalized least squares. *Linear Algebra and its Applications*, 127:449–456, 1990.
27. J. de Leeuw. Centering in multilevel analysis. In B. S. Everitt and D. C. Howell, editors, *Encyclopedia of Statistics in Behavioral Science*, volume 1, pages 247–249. Wiley, New York, 2005.
28. J. de Leeuw and I. G. G. Kreft. Random coefficient models for multilevel analysis. *Journal of Educational Statistics*, 11:57–85, 1986.
29. J. de Leeuw and I. G. G. Kreft. Questioning multilevel models. *Journal of Educational and Behavioral Statistics*, 20:171–190, 1995.
30. J. de Leeuw and I. G. G. Kreft. Software for multilevel analysis. In A. H. Leyland and H. Goldstein, editors, *Multilevel Modelling of Health Statistics*, pages 187–204. Wiley, Chichester, 2001.
31. J. de Leeuw and G. Liu. Augmentation algorithms for mixed model analysis. Preprint 115, UCLA Statistics, Los Angeles, CA, 1993.
32. G. del Pino. The unifying role of iterative generalized least squares in statistical algorithms. *Statistical Science*, 4:394–408, 1989. (with discussion)
33. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the *EM* algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977. (with discussion)
34. A. P. Dempster, D. B. Rubin, and R. K. Tsutakawa. Estimation in covariance components models. *Journal of the American Statistical Association*, 76:341–353, 1981.
35. M. du Toit and S. H. C. du Toit. *Interactive LISREL: User's Guide*. Scientific Software International, Chicago, 2002.
36. W. J. Duncan. Some devices for the solution of large sets of simultaneous linear equations (with an appendix on the reciprocation of partitioned matrices). *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science, 7th Series*, 35:660–670, 1944.
37. J. P. Elhorst and A. S. Zeilstra. Labour force participation rates at the regional and national levels of the European Union: An integrated analysis. *Papers in Regional Science*, forthcoming.
38. T. S. Ferguson. *A Course in Large Sample Theory*. Chapman & Hall, London, 1996.

39. D. A. Freedman. On the so-called "Huber sandwich estimator" and "robust standard errors". Unpublished manuscript, 2006.
40. A. Gelman. Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics*, 48:432–435, 2006.
41. A. Gelman and I. Pardoe. Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics*, 48:241–251, 2006.
42. A. Gelman, D. K. Park, S. Anselobehere, P. N. Price, and L. C. Minnete. Models, assumptions and model checking in ecological regressions. *Journal of the Royal Statistical Society, Series A*, 164:101–118, 2001.
43. A. S. Goldberger. Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association*, 57:369–375, 1962.
44. H. Goldstein. Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73:43–56, 1986.
45. H. Goldstein. Restricted unbiased iterative generalized least-squares estimation. *Biometrika*, 76:622–623, 1989.
46. H. Goldstein. *Multilevel Statistical Models*, 3rd edition. Edward Arnold, London, 2003.
47. H. Goldstein and J. Rasbash. Efficient computational procedures for the estimation of parameters in multilevel models based on iterative generalised least squares. *Computational Statistics & Data Analysis*, 13:63–71, 1992.
48. P. J. Green. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society, Series B*, 46:149–192, 1984. (with discussion)
49. L. Grilli and M. Pratesi. Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs. *Survey Methodology*, 30:93–103, 2004.
50. E. A. Hanushek. Efficient estimates for regressing regression coefficients. *American Statistician*, 28:66–67, 1974.
51. H. O. Hartley and J. N. K. Rao. Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54:93–108, 1967.
52. D. A. Harville. Baysian inference for variance components using only error contrasts. *Biometrika*, 61:383–385, 1974.
53. D. A. Harville. *Matrix Algebra From a Statistician's Perspective*. Springer, New York, 1997.
54. T. Hastie and R. Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society, Series B*, 55:757–796, 1993. (with discussion)
55. D. Hedeker. MIXNO: A computer program for mixed-effects nominal logistic regression. *Journal of Statistical Software*, 4(5):1–92, 1999.
56. D. Hedeker and R. D. Gibbons. MIXOR: A computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine*, 49:157–176, 1996.
57. D. Hedeker and R. D. Gibbons. MIXREG: A computer program for mixed-effects regression analysis with autocorrelated errors. *Computer Methods and Programs in Biomedicine*, 49:229–252, 1997.

58. J. Hemelrijk. Underlining random variables. *Statistica Neerlandica*, 20:1–7, 1966.
59. J. J. Hox. *Multilevel Analysis: Techniques and Applications*. Erlbaum, Mahwah, NJ, 2002.
60. C. Hsiao. *Analysis of Panel Data*, 2nd edition. Cambridge University Press, Cambridge, UK, 2003.
61. J. Z. Huang, C. O. Wu, and L. Zhou. Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, 89:111–128, 2002.
62. C. Jencks, M. Smith, H. Acland, M. J. Bane, D. Cohen, H. Gintis, B. Heyns, and S. Michelson. *Inequality: A Reassessment of the Effect of Family and Schooling in America*. Basic Books, New York, 1972.
63. R. I. Jennrich and M. D. Schluchter. Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 42:805–820, 1986.
64. K. G. Jöreskog, D. Sörbom, S. H. C. du Toit, and M. du Toit. *LISREL 8: New Statistical Features*. Scientific Software International, Chicago, 2001. (3rd printing with revisions)
65. J. Kim and E. W. Frees. Multilevel modeling with correlated effects. *Psychometrika*, forthcoming.
66. M. S. Kovačević and S. N. Rai. A pseudo maximum likelihood approach to multilevel modelling of survey data. *Communications in Statistics—Theory and Methods*, 32:103–121, 2003.
67. I. G. G. Kreft and J. de Leeuw. *Introducing Multilevel Modeling*. Sage, London, 1998.
68. I. G. G. Kreft, J. de Leeuw, and L. S. Aiken. The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, 30: 1–21, 1995.
69. N. M. Laird, N. Lange, and D. Stram. Maximum likelihood computations with repeated measures: Application of the EM algorithm. *Journal of the American Statistical Association*, 82:97–105, 1987.
70. N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38:963–974, 1982.
71. L. I. Langbein. Schools or students: Aggregation problems in the study of student achievement. *Evaluation Studies Review Annual*, 2:270–298, 1977.
72. D. V. Lindley and A. F. M. Smith. Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, 34:1–41, 1972.
73. M. J. Lindstrom and D. M. Bates. Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83:1014–1022, 1988.
74. N. T. Longford. A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, 74:817–827, 1987.
75. N. T. Longford. *VARCL. Software for Variance Component Analysis of Data with Nested Random Effects (Maximum Likelihood)*. Educational Testing Service, Princeton, NJ, 1990.

76. N. T. Longford. *Random Coefficient Models*. Oxford University Press, Oxford, UK, 1993.
77. K. Löwner. Über monotone Matrixfunktionen. *Mathematische Zeitschrift*, 38: 177–216, 1934.
78. C. J. M. Maas and J. J. Hox. The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics & Data Analysis*, 46:427–440, 2004.
79. J. R. Magnus and H. Neudecker. Symmetry, 0–1 matrices and Jacobians: A review. *Econometric Theory*, 2:157–190, 1986.
80. J. R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley, Chichester, 1988.
81. W. M. Mason, G. Y. Wong, and B. Entwisle. Contextual analysis through the multilevel linear model. *Sociological Methodology*, 14:72–103, 1983.
82. P. McCullagh and J. A. Nelder. *Generalized Linear Models*, 2nd edition. Chapman & Hall, London, 1989.
83. E. Meijer and J. Rouwendal. Measuring welfare effects in models with random coefficients. *Journal of Applied Econometrics*, 21:227–244, 2006.
84. B. O. Muthén and A. Satorra. Complex sample data in structural equation modeling. *Sociological Methodology*, 25:267–316, 1995.
85. L. K. Muthén and B. O. Muthén. *Mplus User's Guide*, 4th edition. Muthén & Muthén, Los Angeles, 1998–2006.
86. J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, 1999.
87. D. Pfeffermann. The role of sampling weights when modeling survey data. *International Statistical Review*, 61:317–337, 1993.
88. D. Pfeffermann, C. J. Skinner, D. J. Holmes, H. Goldstein, and J. Rasbash. Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B*, 60:23–56, 1998. (with discussion)
89. J. C. Pinheiro and D. M. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer, New York, 2000.
90. J. C. Pinheiro, D. M. Bates, S. DebRoy, and D. Sarkar. *The nlme Package*, 2006. URL `http://cran.r-project.org`
91. R. F. Potthoff and S. N. Roy. A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 51:313–326, 1964.
92. R. F. Potthoff, M. A. Woodbury, and K. G. Manton. "Equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey weights under superpopulation models. *Journal of the American Statistical Association*, 87:383–396, 1992.
93. R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. URL `http://www.r-project.org`
94. S. Rabe-Hesketh and A. Skrondal. Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society, Series A*, 169:805–827, 2006.

95. S. Rabe-Hesketh, A. Skrondal, and A. Pickles. GLLAMM manual. Working Paper 160, U.C. Berkeley Division of Biostatistics, Berkeley, CA, 2004. (Downloadable from `http://www.bepress.com/ucbbiostat/paper160/`)
96. C. R. Rao. *Linear Statistical Inference and its Applications*, 2nd edition. Wiley, New York, 1973.
97. J. Rasbash, F. Steele, W. J. Browne, and B. Prosser. *A User's Guide to MLwiN. Version 2.0.* Centre for Multilevel Modelling, University of Bristol, Bristol, UK, 2005.
98. S. W. Raudenbush. Reexamining, reaffirming, and improving application of hierarchical models. *Journal of Educational and Behavioral Statistics*, 20:210–220, 1995.
99. S. W. Raudenbush and A. S. Bryk. Empirical Bayes meta-analysis. *Journal of Educational Statistics*, 10:75–98, 1985.
100. S. W. Raudenbush and A. S. Bryk. A hierarchical model for studying school effects. *Sociology of Education*, 59:1–17, 1986.
101. S. W. Raudenbush and A. S. Bryk. *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd edition. Sage, Thousand Oaks, CA, 2002.
102. S. W. Raudenbush, A. S. Bryk, Y. F. Cheong, and R. Congdon. *HLM 6: Hierarchical Linear and Nonlinear Modeling.* Scientific Software International, Chicago, 2004.
103. G. K. Robinson. That BLUP is a good thing: the estimation of random effects. *Statistical Science*, 6:15–51, 1991. (with discussion)
104. W. S. Robinson. Ecological correlations and the behavior of individuals. *Sociological Review*, 15:351–357, 1950.
105. J. Rouwendal and E. Meijer. Preferences for housing, jobs, and commuting: A mixed logit analysis. *Journal of Regional Science*, 41:475–505, 2001.
106. SAS/Stat. *SAS/Stat User's Guide, version 9.1.* SAS Institute, Cary, NC, 2004.
107. M. D. Schluchter. BMDP5V – Unbalanced repeated measures models with structured covariance matrices. Technical Report 86, BMDP Statistical Software, Los Angeles, 1988.
108. C. J. Skinner. Domain means, regression and multivariate analysis. In C. J. Skinner, D. Holt, and T. M. F. Smith, editors, *Analysis of Complex Surveys*, pages 59–87. Wiley, New York, 1989.
109. C. J. Skinner, D. Holt, and T. M. F. Smith, editors. *Analysis of Complex Surveys.* Wiley, New York, 1989.
110. A. Skrondal and S. Rabe-Hesketh. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models.* Chapman & Hall/CRC, Boca Raton, FL, 2004.
111. T. A. B. Snijders and R. J. Bosker. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling.* Sage, London, 1999.
112. D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 64:583–639, 2002. (with discussion)
113. D. J. Spiegelhalter, A. Thomas, N. G. Best, and D. Lunn. *WinBUGS User Manual, Version 1.4.* MRC Biostatistics Unit, Cambridge, UK, 2003.
114. SPSS. *SPSS Advanced Models™ 15.0 Manual.* SPSS, Chicago, 2006.

115. StataCorp. *Stata Statistical Software: Release 9*. Stata Corporation, College Station, TX, 2005.
116. J. L. F. Strenio, H. I. Weisberg, and A. S. Bryk. Empirical Bayes estimation of individual growth curve parameters and their relationship to covariates. *Biometrics*, 39:71–86, 1983.
117. P. A. V. B. Swamy. *Statistical Inference in a Random Coefficient Model*. Springer, New York, 1971.
118. R. L. Tate and Y. Wongbundhit. Random versus nonrandom coefficient models for multilevel analysis. *Journal of Educational Statistics*, 8:103–120, 1983.
119. G. Van Landeghem, P. Onghena, and J. Van Damme. The effect of different forms of centering in hierarchical linear models re-examined. Technical Report 2001-04, Catholic University of Leuven, University Centre for Statistics, Leuven, Belgium, 2001.
120. G. Verbeke and E. Lesaffre. Large sample properties of the maximum likelihood estimators in linear mixed models with misspecified random-effects distributions. Technical Report 1996.1, Catholic University of Leuven, Biostatistical Centre for Clinical Trials, Leuven, 1996.
121. G. Verbeke and E. Lesaffre. The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis*, 23:541–556, 1997.
122. G. Verbeke and G. Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer, New York, 2000.
123. T. Wansbeek and E. Meijer. *Measurement Error and Latent Variables in Econometrics*. North-Holland, Amsterdam, 2000.
124. J. M. Wooldridge. Asymptotic properties of weighted M-estimators for variable probability samples. *Econometrica*, 67:1385–1406, 1999.
125. J. M. Wooldridge. Asymptotic properties of weighted M-estimators for standard stratified samples. *Econometric Theory*, 17:451–470, 2001.
126. J. M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA, 2002.
127. J. M. Wooldridge. Inverse probability weighted M-estimators for sample selection, attrition, and stratification. *Portuguese Economic Journal*, 1:117–139, 2002.
128. R. Xu. Measuring explained variation in linear mixed effects models. *Statistics in Medicine*, 22:3527–3541, 2003.