Universiteit Leiden

MLA

Software for MultiLevel Analysis of Data with Two Levels

User’s Guide for Version 4.1

Frank M. T. A. Busing
Leiden University, Department of Psychology,
P.O. Box 9555, 2300 RB Leiden, The Netherlands.

Erik Meijer
University of Groningen, Department of Econometrics,
P.O. Box 800, 9700 AV Groningen, The Netherlands.

Rien van der Leeden
Leiden University, Department of Psychology,
P.O. Box 9555, 2300 RB Leiden, The Netherlands.

December 2005

Department of Psychology
Faculty of Social and Behavioural Sciences

This document can be referenced as follows:

Busing, F. M. T. A., Meijer, E., & Van der Leeden, R. (2005). *MLA. Software for multilevel analysis of data with two levels. User's guide for version 4.1*. Leiden, The Netherlands: Leiden University, Department of Psychology.

Preface

This manual describes MLA, version 4.1, a computer program developed for multilevel analysis of data with two levels. The first version of MLA was written in 1993–1994, to empirically evaluate the bootstrap (and jackknife) estimators for multilevel analysis that we had developed, and continued to develop, extend, and improve. This original project was supported by a grant (SVO project no. 93713) from the Institute for Educational Research in the Netherlands (SVO). Therefore, we thank SVO for their financial support. In the years following its introduction, MLA has been extended and improved gradually, mainly by implementation of improved and extended resampling estimators that have been the results of our research, but also by implementation of user-requested features or features that we found useful ourselves. The new versions have been accompanied by short "readme" files that briefly described the new features, but a proper new version of the manual has been embarassingly lacking. Here, we finally provide an up-to-date manual.

The MLA program can still be characterized predominantly by our need to have a platform for our research on resampling methods. Nevertheless, from the feedback that we get, we may conclude that it has proven to be useful for a wider audience. We expect that this new manual, and the latest version of the program (4.1) that it describes, enhance its usefulness. The preface of the first version of this manual (Busing, Meijer, & Van der Leeden, 1994) gave the following description of the features of MLA:

> The MLA program can be characterized by four major properties:
>
> - User-friendly interface.
> - Extensive options for simulation, in particular, three options for bootstrapping multilevel models.
> - Simple estimation methods, providing an alternative for the complex iterative estimation procedures that are commonly used to estimate the parameters of multilevel models.
> - A fast algorithm, using the Broyden-Fletcher-Goldfarb-Shanno optimization method to obtain maximum likelihood estimates of all model parameters.

The first point must perhaps be amended a little, in light of the graphical user interfaces common today, although we still think that the command syntax used by MLA is very easy and intuitive.

The MLA program runs as a stand-alone batch program in command windows on personal computers under Windows. It uses simple ASCII text files as input and output. The program is easy to use by means of a number of statements starting with a keyword. Models are specified by simply formulating the model equations.

This manual provides the necessary information for the new user to fit multilevel models with two levels to a hierarchical data set. It is expected that the user has basic knowledge of regression analysis. A brief introduction to multilevel analysis and related concepts is given in the first chapter. References to relevant literature can be found in the text.

The MLA program has been developed by Frank Busing, Erik Meijer, and Rien van der Leeden. Although the research subjects that led us to write MLA are not our main subjects anymore, and thus work on MLA does not have our highest priority anymore as well, we would still appreciate hearing about your experiences using the program and this manual. Please contact us by email: busing@fsw.leidenuniv.nl, e.meijer@eco.rug.nl, or vanderleeden@fsw.leidenuniv.nl.

We would like to thank Jan de Leeuw, Ita Kreft, Wolfgang Langer, and Tom Wansbeek for helpful discussions, comments, references, etc.

Frank M. T. A. Busing
Erik Meijer
Rien van der Leeden

Leiden and Groningen, December 2005

Contents

1 Introduction **1**

1.1 Introduction to multilevel analysis 1

1.2 The position of MLA 5

1.3 Changes since version 1.0b 7

2 Theory **9**

2.1 The general two-level model 9

2.2 Descriptive statistics 10

2.3 Centering 12

2.3.1 Centering in linear regression 12

2.3.2 Centering around the grand mean in a multilevel model 13

2.3.3 Level-2 centering 15

2.4 Ordinary least squares estimators 18

2.4.1 Within-group models 18

2.4.2 One-step OLS (total model) 19

2.4.3 Two-step OLS (total model) 19

2.5 Full Information Maximum Likelihood (FIML) 20

2.6 Restricted Maximum Likelihood (REML) 21

2.7 Residuals 21

2.8 Posterior means 22

2.9 Diagnostics 23

2.10 Simulation 23

2.10.1 The jackknife 25

2.10.2 Jackknife confidence intervals 28

2.10.3 The bootstrap 28

2.10.4 Balanced bootstrap 30

2.10.5 Resampling regression models 30

2.10.6 Resampling multilevel models 31

2.10.7 Bootstrap confidence intervals 33

2.11 Effective sample size and intra-class correlation 37

2.12 Missing data 39

3 Commands **41**

3.1 /TITLE (optional) 41

3.2 /DATA (required) 42

3.2.1 file (required) 42

3.2.2 variables (required) 42

3.2.3 id1 (optional) 42
3.2.4 id2 (required) 42
3.2.5 missing (optional) 43
3.2.6 centering (optional) 43
3.2.7 level-2 centering (optional) 43
3.3 /MODEL (required) 43
3.4 /CONSTRAINTS (optional) 44
3.5 /TECHNICAL (optional) 45
3.5.1 estimation (optional) 45
3.5.2 minimization (optional) 45
3.5.3 reparameterization (optional) 46
3.5.4 warnings (optional) 46
3.5.5 maxiter (optional) 46
3.5.6 convergence (optional) 46
3.5.7 seed (optional) 47
3.5.8 file (optional) 47
3.6 /SIMULATION (optional) 47
3.6.1 kind (required) 47
3.6.2 method (optional) 48
3.6.3 type (optional) 49
3.6.4 balancing (optional) 49
3.6.5 resample (optional) 49
3.6.6 linking (optional) 49
3.6.7 replications (optional) 49
3.6.8 convergence (optional) 49
3.6.9 file (optional) 50
3.7 /INTERVAL (optional) 50
3.7.1 kind (required) 50
3.7.2 alpha (optional) 50
3.7.3 weight (optional) 51
3.7.4 replications (optional) 51
3.7.5 convergence (optional) 51
3.7.6 file (optional) 51
3.8 /PRINT (optional) 51
3.8.1 input (optional) 51
3.8.2 descriptives (optional) 52
3.8.3 random level-1 coefficients (optional) 52
3.8.4 olsquares (optional) 52
3.8.5 residuals (optional) 52
3.8.6 posterior means (optional) 52
3.8.7 diagnostics (optional) 52
3.9 /PLOT (optional) 53
3.9.1 histograms (optional) 53
3.9.2 scatters (optional) 53

4 Worked examples **55**
4.1 Random effects analysis of variance 55
4.2 Random effects analysis of covariance 61
4.3 Repeated measures analysis 62
4.4 Multilevel analysis with bootstrap 65

A Technical appendix **73**
A.1 The model and the likelihood function 74
A.2 Some useful formulas 75
A.3 Computational formulas for the function and gradient 79
A.4 The asymptotic covariance matrix of the estimators 84
A.5 Reparametrization 93

References **97**

Chapter 1

Introduction

1.1 Introduction to multilevel analysis

Multilevel analysis comprises a set of techniques that explicitly take into account the hierarchical structure in the data. In this section, a brief introduction to the underlying ideas of multilevel analysis is given. Several relevant topics, such as hierachical data structures, intra-class correlation, the formulation of a multilevel model, and the estimation of the model parameters are discussed. This introduction does not contain formulas. Chapter 2 will discuss the main formulas, and the Technical Appendix will give supplementary mathematical details.

Hierarchical data

Hierarchically structured data arise in a variety of research areas. Such data are characterized by so-called “nested” membership relations among the units of observation. Classical examples of hierarchically structured data are found in educational research where, for instance, students are nested within classes and classes are nested within schools. But, in many other instances in the social and behavioral sciences, as well as in many other fields of science, data are also hierarchically structured. For instance, in clinical psychology, clients can be nested within therapy groups, people can be nested within families, and so forth. A sociological example is given by a study concerning employees nested within industries.

It should be noted that nested structures naturally arise where explicit hierarchical sampling schemes are used. This is often the case in large scale educational research where, for instance, a set of schools is sampled first, followed by the sampling of a set of students within these schools. However, there are many other cases where data are not explicitly sampled in that way, but where it appears to be a fruitful approach to treat them as having a hierarchical structure. For instance, in a medical study one could consider it to be important that patients can be viewed as nested within general practitioners. Apart from this, there are several types of data for which it proves to be very useful to apply the concept of hierarchy, because it makes their analysis more easy and transparent. One example is the hierarchical treatment of repeated measures data, where measurements at different points in time are considered nested within individuals. Another example is the analysis of data from meta analysis, where, say, p-values can be treated as being nested within studies, providing a (partial) solution for the problem of comparing apples with oranges.

With hierarchical data, it is common to have information obtained from the different *levels* in the hierarchy. For instance, one has variables describing the individual students, but also variables describing their schools. When analyzing such data one has to decide in what way the hierarchical structure of the data is taken into account. Obviously, the easiest approach is simply ignoring the structure and analyzing the data at the student level, leaving all school information for what it is. Generally, however, one's intention will be to use all information in the data, and use it correctly. Thus, if one is also interested in school differences and in their possible interaction with effects measured at the student level, one has to solve the "unit-of-analysis" problem. This means that one has to decide whether to analyze the data at the student level, incorporating disaggregated variables from the school level, or to analyze the data at the school level, incorporating aggregated variables from the student level. Unfortunately, both of these strategies are subject to serious disadvantages (De Leeuw, 2005b). Hence, traditional "single level" analyses fail in the presence of nested data.

Intra-class dependency

The basic problem with hierarchical data is that group membership may cause *intra-class dependency*: People from the same group are more alike than people from different groups. The reason for this phenomenon is that people within a group share the same environment, have the same leader, experiences, and so forth. In other words, people within the same group have the same score on a number of variables, most of them never measured and thus omitted from any possible model. Hence, if we fit a (common, single level) model to such data, intra-class dependency has an effect on the error terms. It causes the error terms to be correlated. The result is that the usual assumption of independent observations is violated if the nested structure of the data is ignored. The degree of intra-class dependency is reflected in the *intra-class correlation*. Obviously, this idea of intra-class dependency applies to every hierarchical data set. Their intra-class correlations, however, may differ substantially.

Multilevel models

For the analysis of hierarchical data, hierarchical models, or "multilevel" models have been developed. Such models can be conceived as linear regression models, specified separately for each level of the hierarchy, but statistically connected. Since each level of the hierarchy has its own regression equation, predictor variables measured at either level can be included in the appropriate level model.

Because hierarchical data structures frequently arise in social and behavioral science research, but also in many other scientific areas, the application and development of multilevel analysis has in the last decade drawn a lot of attention from numerous researchers. Below, a brief introduction of some relevant topics concerning multilevel models will be given. A more comprehensive introduction of these topics is given by Kreft and Van der Leeden (1994). For extensive discussions on theory and application of multilevel analysis, we refer to the literature. An incomplete list of textbooks is Kreft and De Leeuw (1998), Snijders and Bosker (1999), Hox (2002), Raudenbush and Bryk (2002), Goldstein (2003), and Langer (2004). The latter uses `MLA` in his empirical examples.

Small example

A small, imaginary, example from education may clarify what is meant by a multilevel model. Suppose we have data of students nested within schools, and we want to predict the score on a math test from the amount of time spend on doing math homework. Furthermore, we expect smaller schools to be more effective than larger ones, so we collect the school size as another variable. Clearly, at the student level, 'math' is the dependent variable, and 'homework' is the predictor variable. At the school level, ‘size’ is the predictor variable. Now the multilevel model for this example, in this case a two-level model, is specified as follows. At the student level, Level-1, for each school a regression model is formulated with ‘math’ as the dependent variable and ‘homework’ as the predictor. This reflects the intra-class dependency of the observations (the students) within each school: All models contain the same variables, but we expect them to yield different intercept and slope estimates within each school. At the school level, Level-2, a regression model is formulated in which the intercepts and slopes of the first-level models are dependent variables, predicted by the second-level variable ‘size’. This reflects the possible effect of school size on school effectiveness: School size may influence the estimated relationship between ‘math’ and ‘homework’.

At first glance, the model presented above seems to lead to a hierarchically structured regression procedure, which proceeds in two steps: First, the models for all schools are estimated, and then the intercept and slope estimates are used as the dependent variables in the Level-2 model, which is then estimated. Although such procedures have been proposed in the past, this is not what will be discussed here under the heading of multilevel models, because there is no statistical connection between the Level-1 and Level-2 models. In multilevel models, separate regression equations for each level are only formulated because they facilitate insight and understanding. The statistical linkage of both levels is created by the Level-2 model which states that Level-1 regression coefficients—intercepts and slopes—are treated as *random variables* at the second level. The Level-2 model models intercept and slope estimates as a mean value over all schools plus a school-specific deviation or residual. It follows that we are not primarily looking for intercept and slope estimates for each separate school, but for their means and variances (and their covariance) over all schools. In this way, just as students are considered a sample from a population of students, schools are considered a sample from a population of schools.

There are several reasons why it may be useful to consider the school-specific coefficients as random. First, the schools in the data set are usually a random sample from the “population” of schools, and scientists are usually interested in the population, rather than the specific data set. Second, with a model that explains part of the variation in the random coefficients, the effect of the school-level variables on the student-level relationships can be assessed, and, in particular, the model can give guidance to schools that want to improve their effectiveness. Third, the relationships between the outcome variable and the student-level predictors become clearer: Between-school variation that may blur these relationships is accounted for, and consequently, the estimates of the average coefficients are more precise.

School-specific estimates of intercept and slope can, however, be obtained. This will be discussed below under the heading of Random Level-1 coefficients.

Cross-level interaction

If a school-level predictor variable like ‘size’ is added to the Level-2 model in our imaginary example, means and variances change to conditional means and variances. It means that part of the variance of intercepts and slopes among schools is explained by ‘size’. The contribution of this school-level variable introduces a term to the model that specifies a relationship between both levels: The relationship between ‘size’ at the school level and the slope coefficient for each separate school, which is part of the model at the student level. As was said above, this term refers to the expected influence of ‘size’ on the regression of ‘math’ on ‘homework’. In the terminology of multilevel analysis this term is called a *cross-level interaction*. For some researchers, this interaction term provides the main attraction to multilevel analysis. It is the cross-level interaction parameter that leads to the interpretation of “slopes-as-outcomes” (cf. Aitkin & Longford, 1986).

The number of levels

Theoretically, we can model as many levels as we know the hierarchy has, or as we think it will have. In practice, however, most applications of multilevel analysis concern problems with two or three levels. Data sets with more than three levels are rare. In fact, a majority of applications just concerns two-level data and can be viewed as “within-and-between-analysis” problems. It should be noted that models with more than three levels show a rapid increase in complexity, especially where interpretation is concerned. If such models are necessary, they should be limited to rather simple cases, that is, to cases with only a few predictor variables.

Random Level-1 coefficients

In multilevel modeling, we are usually not looking for estimates of the regression coefficients within each separate group, but for their variances and covariances. However, there can be circumstances in which we still want to obtain the “best” estimates for these coefficients, also called *random Level-1 coefficients*. Such questions may arise, for example, in education when schools are to be ranked in terms of effectiveness, using their estimated slope coefficients (Kreft & De Leeuw, 1991). The first thing that comes to mind is to simply estimate them by a separate (OLS) regression for each school. However, this procedure has the serious disadvantage that the coefficients will not be estimated with the same precision for each school. For instance, in one school, we could have, say, 45 students, whereas in another school we only have 7 students. This will definitely influence the accuracy of results.

Within the framework of multilevel analysis there is a way to obtain best estimates of these coefficients by a method called *shrinkage* estimation. The underlying idea of this estimation is that there are basically two sources of information: the estimates from each group separately and the estimates that could be obtained from the total sample, ignoring any grouping. Shrinkage estimation consists of a weigted combination of these two sources. The more reliable the estimates are within the separate groups, the more weight is put on them. Vice versa, the less reliable these estimates are, that is, the less precise, the more weight is put on the estimates obtained from the total sample. The result is that estimates are “shrunken” towards the mean of the estimates over all groups. The amount of shrinkage depends on the reliability of the estimates from the separate groups. The less precise the estimates are, the more they are “shrunken” towards the mean over all groups. Technically, the shrunken estimators are the expectations of the (random) coefficients given the parameter estimates and the data of all groups.

Estimation

Fitting a multilevel model amounts to fitting one combined model, instead of separate models for each level. It is the translation of the idea that, although separate models for each level may be formulated, they are statistically connected, as was mentioned above. The combined model contains all relevant parameters. In the next chapter, we will further clarify this subject.

Combined models, or multilevel models, can be viewed as special cases of the general mixed linear model (cf. Harville, 1977). Such models are characterized by a set of fixed and a set of random regression coefficients. The parameters that have to be estimated are the fixed coefficients and the variances and covariances of the random coefficients and random error terms. The fixed coefficients are informally called *fixed parameters* and the variances and covariances of the random coefficients and random error terms are informally called *random parameters*, although all these parameters are technically nonrandom. They are the parameters associated with the fixed and random parts of the model, respectively.

To obtain estimates for the parameters, several estimation procedures have been proposed. These procedures are all versions, in one way or another, of full information (FIML) or restricted maximum likelihood (REML). FIML and REML estimators have several attractive properties, such as consistency and efficiency. A drawback of both approaches, however, is their relative complexity. Generally, parameter estimates must be obtained iteratively and serious computational difficulties may arise during such processes.

Software

The flourishing of models and techniques for analyzing hierarchical data has been stimulated by the software widely available for estimating multilevel models. Besides dedicated multilevel modeling software like MLA, HLM (Raudenbush, Bryk, Cheong, & Congdon, 2004), and MLwiN (Rasbash, Steele, Browne, & Prosser, 2004), software on structural equation modeling like LISREL, EQS, M*plus*, and Mx also contain options for multilevel modeling, with varying degrees of generality. Of these, Mx can be freely downloaded from http://www.vipbg.vcu.edu/mxgui/ General purpose statistical packages like SAS and SPSS, and (statistical) matrix programming environments like S-plus and R also offer procedures for multilevel analysis. These typically have names like 'MIXED', 'GLMM' (generalized linear mixed modeling), or 'NLME' (nonlinear mixed estimation). The R program can also be freely downloaded (cran.r-project.org).

The website of the Centre for Multilevel Modelling (http://multilevel.ioc.ac.uk) currently (August 2005) contains the following list of programs that can be used for multilevel modeling: aML, EGRET, GENSTAT, HLM, LIMDEP, LISREL, MIXREG, MLwiN, Mplus, R, SAS, S-Plus, SPSS, STATA, SYSTAT, WINBUGS. Remarkably, they do not include MLA.

Terminology

In the literature multilevel models are referred to under various names. One may find the terms *multilevel mixed effects models* (Goldstein, 1986), *multilevel linear models* (Mason, Wong, & Entwisle, 1983), *hierarchical linear models* (Raudenbush & Bryk, 2002), *random coefficient regression models* (De Leeuw & Kreft, 1986), *random parameter models* (Aitkin & Longford, 1986), *contextual-effects models* (Blalock, 1984), *full contextual models* (Kreft & Van der Leeden, 1994), and *variance components models* (Aitkin & Longford, 1986). Although there are minor differences, all these models are basically the same. In one way or another they are versions of the multilevel model discussed here, or straightforward extensions thereof.

1.2 The position of MLA

This manual describes the use and capabilities of MLA. This program has been developed to analyze data with a two-level hierarchical structure. In this section we will explain the features that distinguish MLA from other programs mentioned above. In other words, we are concerned with the question: What is special about MLA? It should be noted that some of the arguments given here were more pressing when we wrote the first version of the program (1993–1994) than they are now, because some of the facilities of MLA have also been adopted by other packages, at least to some extent.

Simulation options

Much research concerning multilevel analysis has been directed towards the extension and refinement of multilevel theory, including the development of multilevel software, and to applications in other domains than educational research. At the same time, however, several relevant questions of a statistical nature concerning this development are still not answered fully satisfactorily. One major problem is that estimates of parameters and standard errors, as well as hypothesis tests based on them, rely on large sample properties of the estimates. Unfortunately, little is known about the behavior of the estimates when sample size is small (Raudenbush, 1988). An additional problem is that it is usually assumed that the error terms are normally distributed. In practice, this assumption will often be violated, which has other undesirable consequences for using standard error estimates for hypothesis testing and construction of confidence intervals.

Fortunately, there is an increasing number of simulation studies available, which give insight into the quality of estimates of parameters and standard errors under various conditions (Busing, 1993; Van der Leeden & Busing, 1994; Kreft, 1996). Concerning empirical data sets, however, we think that extensive simulation options, in particular options for bootstrapping, are a very useful addition to a program for multilevel analysis.

Therefore, four different simulation methods are implemented in MLA:

1. A bootstrap method that uses the estimated parameters as “true values” of the parameters of a multivariate normal distribution from which new outcome variables are drawn. This method is called the *parametric bootstrap*.

2. A bootstrap method that uses the observed values of outcome and predictor variables for resampling. Thus, whole cases are resampled. Therefore, this is called the *cases bootstrap*.

3. A bootstrap method that uses estimates of the error terms at both levels for resampling. In contrast with the cases bootstrap, this method leaves the regression design unaffected. This is called *error bootstrap* or *residual bootstrap*. Because the error terms at both levels must be estimated in order to be resampled, we need the estimates for the separate Level-1 models (random Level-1 coefficients). As was explained earlier, there are multiple choices for these coefficients. These choices account for additional options that can be used when applying the error bootstrap.

4. The *jackknife*. With this method, one entire group is deleted for each resample. There are as many resamples as there are groups.

Depending on the type of simulation used in MLA and depending on the nature of the data, the user can decide to resample both levels in the data, or only the first or the second level. This feature may be useful, for instance, in analyzing repeated measures data.

Alternative simple estimation methods

Usually, complicated iterative estimation procedures are used to estimate the parameters of multilevel models. From a theoretical and technical point of view, these procedures provide the best estimates that can be obtained. However, in practice, some of the algorithms used may be rather slow under certain conditions. In other cases serious computational difficulties may arise that are not easy to overcome. De Leeuw and Kreft (1995) discuss alternative estimation procedures for both fixed and random parameters in multilevel models that are non-iterative and relatively easy to implement. Moreover, in certain cases the quality of the parameter estimates is rather good. Hence, one could question the real gain of the complicated iterative procedures over these simpler alternatives. Therefore, in MLA, we have implemented a one-step and a two-step OLS procedure. A simple WLS procedure has still to be implemented.

Simple procedures can always be used as an addition to complex ones, and vice versa. Their results can always be compared with the results of the iterative methods. It depends on the data which estimation procedures are to be preferred (De Leeuw & Kreft, 1995; Kreft, 1996).

Fast Maximum Likelihood algorithm

To maximize the likelihood function, the MLA program uses the *Broyden-Fletcher-Goldfarb-Shanno* (BFGS) method (e.g., Nocedal & Wright, 1999). This is a fast and stable method to optimize arbitrary functions. It requires that the function and the gradient (the vector of first derivatives) of the function with respect to the parameters be programmed. It minimizes the function with respect to both fixed and random parameters simultaneously. As such, it resembles most the algorithm used by VARCL (Longford, 1990), although the BFGS method does not compute the inverse of the information matrix at each iteration. The algorithms of MLwinN and HLM alternately update the fixed and the random parameters.

1.3 Changes since version 1.0b

The first publicly available version of MLA was version 1.0b, released in December 1994. This was accompanied by an extensive manual (Busing et al., 1994). In the period 1995–2005, a few updates of MLA have been made and released. These contained mainly some additional options, but also a few improvements and bug fixes. These updates have not been accompanied by updates of the complete manual. Instead, the changes were documented only very briefly in the readme files included in the MLA distribution. Consequently, the current manual is the first update of the full manual since the original manual. Therefore, we briefly list here the changes between MLA version 1.0b and MLA version 4.1, as reflected in the changes in the manuals.

- Two types of centering of variables before the analysis have been included. See section 2.3.

- Restricted maximum likelihood (REML) estimators have been added. These tend to give less biased estimators of the random parameters. See section 2.6.

- Improved grouped jackknife estimators have been implemented. These are more efficient than the grouped jackknife estimators that were implemented in MLA 1.0b. The corresponding jackknife variance estimators are much better than the original ones. See section 2.10.1.

- Balanced bootstrapping has been added for the cases and residual bootstrap. This may be slightly more efficient statistically, although it is computationally a bit more complicated. See section 2.10.3.

- For the residual bootstrap, linked resampling is added, which gives more robust estimates when the Level-1 and Level-2 errors are dependent, although it is less efficient when they are independent (as assumed in the model specification). See section 2.10.3.

- For all bootstrap methods, various types of bootstrap confidence intervals have been added. Some of these are typically (much) better for the random parameters than the ordinary "estimate ± 2 s.e." confidence intervals, because they take the skewness of the distribution of the estimators into account. See section 2.10.3.

- The `/OUTPUT` command has been replaced by the `/PRINT` command, and a few options have been changed. See section 3.8.

- The `/PLOT` has been added, which offers some rough plotting options. See section 3.9.

Chapter 2

Theory

In this chapter, the theoretical background of the general two-level model used in MLA will be discussed. It gives the relevant formulas of the model equations and discusses the assumptions underlying the model and some model specification issues. Furthermore, it briefly discusses the elements of the output of MLA: the estimators, standard errors, confidence intervals, and model fit statistics that are implemented in the program, and their statistical properties. A lot of additional output can be (optionally) requested by the user, such as descriptive statistics, residuals, and diagnostic statistics. This is also explained here.

2.1 The general two-level model

In MLA, the following general two-level model is implemented. Suppose data are obtained from N individuals nested within J groups, with group j containing N_j individuals. Now, for group j ($j = 1, \ldots, J$), y_j is a vector containing values on an outcome variable, X_j is an $N_j \times q$ matrix with fixed, explanatory variables (usually including a constant), β_j is a vector of regression coefficients, and ε_j is a vector with random error terms (vectors and matrices of appropriate dimensions). Then, for each group j, the Level-1 or within-group model can be written as

$$y_j = X_j\beta_j + \varepsilon_j. \tag{2.1}$$

The Level-2 or between-group model can be written as

$$\beta_j = W_j\gamma + u_j, \tag{2.2}$$

where W_j is a $q \times p$ matrix with explanatory variables (usually including a constant) obtained at the group level, γ is a vector containing fixed coefficients and u_j is a vector with error terms. Equation (2.2) clearly illustrates the "slopes-as-outcomes" interpretation, because it gives expresses the coefficients in β_j as outcome variables in a separate Level-2 model. However, substition of Equation (2.2) into Equation (2.1) gives the "total" model equation

$$y_j = X_jW_j\gamma + X_ju_j + \varepsilon_j. \tag{2.3}$$

This is a *mixed linear model* (Harville, 1977) of the form

$$y_j = X_j^*\gamma + Z_ju_j + \varepsilon_j, \tag{2.4}$$

in which $X_j^* = X_jW_j$ and $Z_j = X_j$. Several authors use different notations for the models presented in this chapter and in subsequent chapters. We find the separate model equations (2.1) and (2.2) for the two levels most useful for interpretation of the model and its estimates, and the program input is therefore based on them (see chapter 3). For theoretical purposes, we find the form (2.4) most useful, where usually X_j^* will be simply written as X_j. Therefore, in the following both representations will be used where appropriate, and it will be clear from the context which form is used. For now, we will proceed with the form (2.4).

Generally, it is assumed that $\varepsilon_j \sim \mathcal{N}(0, \sigma_\varepsilon^2 I_{N_j})$ and $u_j \sim \mathcal{N}(0, \Theta)$, where σ_ε^2, the variance of the Level-1 error term, is an unknown (scalar) parameter, and Θ, the covariance matrix of the Level-2 error terms, is a (symmetric) matrix of unknown parameters. The covariance matrix V_j of y_j conditional on X_j and Z_j, that is, the matrix containing the variances and covariances of the random part $Z_j u_j + \varepsilon_j$ in Equation (2.4) conditional on Z_j, is expressed as

$$V_j = Z_j \Theta Z_j' + \sigma_\varepsilon^2 I_{N_j}. \tag{2.5}$$

A model for the complete data follows straightforwardly from stacking the J groups' models in Equation (2.4). Its equation is

$$\begin{pmatrix} y_1 \\ \vdots \\ y_J \end{pmatrix} = \begin{pmatrix} X_1 \\ \vdots \\ X_J \end{pmatrix} \gamma + \begin{pmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Z_J \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_J \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_J \end{pmatrix},$$

or

$$y = X\gamma + Zu + \varepsilon. \tag{2.6}$$

The covariance matrix of the complete data, conditional on X and Z, is

$$\begin{aligned} V &= \begin{pmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Z_J \end{pmatrix} \begin{pmatrix} \Theta & 0 & \cdots & 0 \\ 0 & \Theta & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Theta \end{pmatrix} \begin{pmatrix} Z_1' & 0 & \cdots & 0 \\ 0 & Z_2' & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Z_J' \end{pmatrix} + \sigma_\varepsilon^2 I_N \\ &= \begin{pmatrix} V_1 & 0 & \cdots & 0 \\ 0 & V_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & V_J \end{pmatrix}. \end{aligned}$$

The parameters of the model that have to be estimated are the fixed coefficients (elements of the vector γ), the covariance matrix Θ of the random coefficients, and the variance σ_ε^2 of the errors. The elements of γ are called the *fixed parameters*, and σ_ε^2 and the elements of Θ are jointly called the *random parameters*. In the following, formulas are presented for the various parts of the output of MLA. The order of this chapter is similar to the order of the output of MLA, as will become clear later on.

2.2 Descriptive statistics

When requested, MLA produces the following descriptive statistics: mean, standard deviation, variance, skewness, and kurtosis. Any statistical package will produce these statistics as well.

Before looking at the other output, it may be useful to inspect these statistics. Their formulas are:

$$\begin{aligned}
&\text{mean:} && \widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i,\\
&\text{standard deviation:} && \widehat{\sigma} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \widehat{\mu})^2},\\
&\text{variance:} && \widehat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \widehat{\mu})^2,\\
&\text{skewness:} && \widehat{\mu}_3 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i - \widehat{\mu}}{\widehat{\sigma}}\right)^3,\\
&\text{kurtosis:} && \widehat{\mu}_4 = \left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i - \widehat{\mu}}{\widehat{\sigma}}\right)^4\right] - 3,
\end{aligned}$$

where X_i is the measurement of individual i on a typical variable X and N is the total sample size.

Another descriptive statistic that is provided is the *Kolmogorov-Smirnov Z* statistic. This is a measure of deviation from the normal distribution. It tests whether the observed variable has a normal distribution. It is defined as the maximum distance between the estimated (empirical) cumulative distribution function and the best-fitting cumulative normal distribution function. It is computed as follows (Stephens, 1974). First, sort the values of a given variable X, such that X_1 is the smallest value and X_N is the largest. Then compute $w_i = (X_i - \widehat{\mu})/\widehat{\sigma}$, $i = 1, \ldots, N$, and $z_i = \Phi(w_i)$, where Φ is the cumulative distribution function of the standard normal distribution. Now, Kolmogorov-Smirnov's Z is defined as

$$Z = \max_{1 \le i \le N}\left(\max\left\{z_i - \frac{i-1}{N}, \frac{i}{N} - z_i\right\}\right).$$

The (asymptotic) distribution of Z was derived by Durbin (1973), but this is very complicated and, more importantly, this derivation does not apply to multilevel data, because of the intra-class dependency. Therefore, in `MLA`, a simpler approach is used, where p-values are reported that are based on the assumption that the normal distribution is completely specified beforehand (not estimated). This is also not entirely correct, because μ and σ are estimated, but it is sufficient for descriptive (exploratory) purposes. The formula used (cf. Mood, Graybill, & Boes, 1974, p. 509) is

$$\Pr(Z) = 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2NZ^2}, \tag{2.7}$$

where the series is truncated after convergence of the sum. Note, however, that normality of the observed variables is by no means essential for the validity of the multilevel model specification, even if `FIML` is used. The latter only assumes that the distribution of the error terms is normal, which implies that the conditional distribution of y given X and Z is normal. This leaves considerable freedom for the marginal distributions of the observed variables.

Finally, `MLA` computes a number of quantiles of the variables. These are the minimum and maximum and the 5th, 25th, 50th, 75th, and 95th percentile. The k-th percentile is defined as the value such that $k\%$ of the observations on a variable are smaller than this value and $(100 - k)\%$ larger. The 25th and 75th percentile are the first and third quartile and the 50th percentile is the median.

2.3 Centering

In social science data, the 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 *Level-2 centering*, or *within-groups centering*, where the sample average of only the observations within the same Level-2 unit is subtracted. Both forms are implemented in `MLA` as options. Here, we will discuss the basic issues involved. For a more extensive analysis, see Kreft, De Leeuw, and Aiken (1995), Van Landeghem, Onghena, and Van Damme (2001), De Leeuw (2005a), and the references therein.

2.3.1 Centering in linear regression

Let us start by considering a simple bivariate regression model:

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, \tag{2.8}$$

where β_1 and β_2 are fixed coefficients and the ε_i are i.i.d. random errors with zero mean and variance σ^2_ε, and the other notation is obvious. Suppose now that we center the explanatory variable x. Then, in the model specification, x_i is replaced by $x^*_i = x_i - \bar{x}$, where $\bar{x}$ is the sample average of x. This leads to the model

$$y_i = \beta^*_1 + \beta^*_2 x^*_i + \varepsilon^*_i, \tag{2.9}$$

with $\mathrm{Var}(\varepsilon^*_i) = \sigma^2_{\varepsilon^*}$. It is easy to see that, without restrictions on the parameters, the models (2.8) and (2.9) are equivalent: if (2.8) and its parameters are given, we can write

$$\begin{aligned} y_i &= \beta_1 + \beta_2 x_i + \varepsilon_i \\ &= \beta_1 + \beta_2(x_i - \bar{x} + \bar{x}) + \varepsilon_i \\ &= (\beta_1 + \beta_2 \bar{x}) + \beta_2 x^*_i + \varepsilon_i. \end{aligned}$$

Thus, by noting that $\bar{x}$ is a given constant, because we condition on x, we find $\beta^*_2 = \beta_2$, $\beta^*_1 = \beta_1 + \beta_2 \bar{x}$, and $\sigma^2_{\varepsilon^*} = \sigma^2_\varepsilon$ and the model (2.9) satisfies all the usual assumptions. Analogously, by starting from (2.9), we obtain (2.8) with $\beta_2 = \beta^*_2$, $\beta_1 = \beta^*_1 - \beta^*_2 \bar{x}$, and $\sigma^2_\varepsilon = \sigma^2_{\varepsilon^*}$.

Although the models are equivalent, one form may be easier to interpret than the other. The difference between them is the intercept parameter. In (2.8), β_1 is the mean of y for $x = 0$. If $x = 0$ is not a relevant value, then the intercept does not have a useful interpretation. For example, it is not relevant to know the mean intellectual achievement (y) of someone with IQ score (x) equal to zero. In (2.9), β_1 is the mean of y for $x = \bar{x}$. Clearly, the mean intellectual achievement of someone with average IQ is an easily interpretable and highly relevant number. On the other hand, if x is gender, coded 0 for males and 1 for females, then it is not very interesting to know what the mean achievement of someone with “average gender” is, whereas the mean achievement of males ($x = 0$) is easily interpretable and highly relevant. So it depends on the meaning and coding of the variables whether centering is interpretationally useful.

Instead of, or in addition to, the centering of the explanatory variable x, we can also center the outcome variable y, giving $y_i^* = y_i - \bar{y}$. Starting with (2.8), we obtain

$$\begin{aligned} y_i^* &= (-\beta_2 \bar{x}) + \beta_2 x_i + (\varepsilon_i - \bar{\varepsilon}) && (2.10) \\ &= \beta_2 x_i^* + (\varepsilon_i - \bar{\varepsilon}). && (2.11) \end{aligned}$$

In the first of these equations, the "new" intercept is $-\beta_2 \bar{x}$, which seems hard to interpret generally, whereas in the second, the intercept has disappeared. Starting from (2.8), the zero intercept in (2.11) is just a special case, so nothing to worry about. However, starting with (2.11) as the model specification, i.e., without intercept, (2.8) cannot be written in the same form anymore. The intercept has to be introduced again. This is not a problem, but a logical consequence of the choice with respect to centering. In this situation, this is obvious, but we will see below that similar specification issues occur in multilevel models, where they are not so obvious anymore.

The error term in (2.10) and (2.11) is $(\varepsilon_i - \bar{\varepsilon})$. The presence of the mean of the errors means that the centered error terms are not independent anymore. The covariance matrix of the error terms is not diagonal anymore. Moreover, it is singular. Although this seems like an important problem with centering of the outcome variable, it is actually not. The OLS estimator of β_2 is still efficient and equivalent to the OLS estimator based on (2.8) and the OLS standard errors are also equivalent.

2.3.2 Centering around the grand mean in a multilevel model

Now let us look at the following two-level random coefficients model:

$$\begin{aligned} y_{ij} &= \beta_{1j} + \beta_{2j} x_{ij} + \varepsilon_{ij} && (2.12a) \\ \beta_{1j} &= \gamma_1 + \gamma_2 w_{1j} + u_{1j} && (2.12b) \\ \beta_{2j} &= \gamma_3 + \gamma_4 w_{2j} + u_{2j}, && (2.12c) \end{aligned}$$

where β_{1j} and β_{2j} are random coefficients, the γ's are fixed coefficients, w_{1j} and w_{2j} are two Level-2 explanatory variables, and u_{1j} and u_{2j} are random residuals with zero means and covariance matrix Θ. We will now try to rewrite this model in a form with x_{ij} centered around the grand mean:

$$\begin{aligned} y_{ij} &= (\beta_{1j} + \beta_{2j}\bar{x}) + \beta_{2j}(x_{ij} - \bar{x}) + \varepsilon_{ij} \\ &= \beta_{1j}^* + \beta_{2j} x_{ij}^* + \varepsilon_{ij} \\ \beta_{1j}^* &= (\gamma_1 + \gamma_3 \bar{x}) + \gamma_2 w_{1j} + (\gamma_4 \bar{x}) w_{2j} + (u_{1j} + \bar{x} u_{2j}) \\ &= \gamma_1^* + \gamma_2 w_{1j} + \gamma_5 w_{2j} + u_{1j}^*, \end{aligned}$$

where $\gamma_5 = \gamma_4 \bar{x}$ and β_{2j} is the same as in (2.12c). The variance of u_{1j}^* is

$$\Theta_{11}^* = \Theta_{11} + 2\bar{x}\Theta_{21} + \bar{x}^2 \Theta_{22}$$

and the covariance of u_{1j}^* and u_{2j} is

$$\Theta_{21}^* = \Theta_{21} + \bar{x}\Theta_{22}.$$

This shows that after the centering, a variable that was included in the equation for β_{2j} but was not included in the equation for β_{1j} now is included in the equation for β^*_{1j}. Furthermore, even if Θ_{21} is restricted to be zero, Θ^*_{21} is nonzero. This illustrates two practical points for the specification. The first is that, as noted above, the origin of many variables is arbitrary. We now see that by a simple change of origin a variable that was not included in the equation for the random intercept becomes part of this equation. The conclusion must thus be that all explanatory variables that are included in one or more of the equations for the random slopes must also be included in the equation for the random intercept. From (2.3), this can also be interpreted as that for every interaction effect, the main effect must also be included. This is a very natural requirement that is also common in the analysis of variance.

The second practical point is that by a simple change of origin of an explanatory variable, a zero covariance between the residuals of the random intercept and a random slope becomes nonzero. Therefore, the covariances of the residual of the random intercept with the rresiduals of the random slopes should not be restricted to zero (or any other fixed value). This last point was also stressed by Snijders and Bosker (1999, 69–70).

If we take these practical points into account, then it follows that centering a Level-1 explanatory variable around the grand mean leads to an equivalent model and thus, the choice whether or not to center can be confidently based on issues of interpretation.

Let us now study what happens when we center a Level-2 variable. In the light of the previous results, we consider the following two-level random coefficients model:

$$y_{ij} = \beta_{1j} + \beta_{2j}x_{ij} + \varepsilon_{ij} \tag{2.13a}$$

$$\beta_{1j} = \gamma_1 + \gamma_2 w_j + u_{1j} \tag{2.13b}$$

$$\beta_{2j} = \gamma_3 + \gamma_4 w_j + u_{2j}, \tag{2.13c}$$

i.e., (2.12), but with the same explanatory variable in both Level-2 equations. We could add more regressors to the intercept equation, but this does not affect the discussion. We want to know what happens if we center w_j. Because the equation for y_{ij} does not explicitly contain w_j, we focus attention to the Level-2 equations. The intercept equation can be rewritten as

$$\begin{aligned}\beta_{1j} &= (\gamma_1 + \gamma_2\bar{w}) + \gamma_2(w_j - \bar{w}) + u_{1j}\\ &= \gamma^*_1 + \gamma_2 w^*_j + u_{1j},\end{aligned}$$

where $w^*_j = w_j - \bar{w}$ is the centered Level-2 regressor and $\gamma^*_1 = \gamma_1 + \gamma_2\bar{w}$ is the new intercept. The analysis is completely analogous to the analysis of the effect of centering in an ordinary regression model: if there are no restrictions on γ_1, then the centered and uncentered models are equivalent, and the advantages and disadvantages of centering are analogous to those for the ordinary regression model. It is immediately clear that this carries over to the slope equation as well.

Note that we have not properly defined $\bar{w}$. There are two sensible definitions, one explicitly recognizing that w is a Level-2 variable and one that takes the simple average over all Level-1 units:

$$\bar{w}_{(2)} = \frac{1}{J}\sum_{j=1}^{J} w_j$$

$$\bar{w}_{(1)} = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{N_j} w_j = \frac{1}{N}\sum_{j=1}^{J} N_j w_j = \frac{1}{J}\sum_{j=1}^{J}\frac{JN_j}{N} w_j.$$

For balanced data, i.e., if $N_j = N/J$ for all j, then these two definitions revert to the same average. Otherwise, $\bar{w}_{(1)}$ attaches proportionally more weight to Level-2 units with more observations. Note that these are not necessarily "larger" units in the population. Just as centered and uncentered models are equivalent, the models with different forms of centering are statistically equivalent, and preference for one or another are purely interpretational. MLA uses $\bar{w}_{(1)}$ as a basis for grand mean centering, which conforms to centering of a Level-1 variable. Users who prefer centering around $\bar{w}_{(2)}$ should compute the centered variable outside of MLA and include this in the data file; within MLA this variable should then not be centered, of course.

Centering of the outcome variable is more complicated. Consider the model (2.13). Then

$$\begin{aligned} y^*_{ij} &= y_{ij} - \bar{y} \\ &= \beta_{1j} + \beta_{2j}x_{ij} + \varepsilon_{ij} - \bar{\beta}_1 - \overline{\beta_2 x} - \bar{\varepsilon} \\ &= (\beta_{1j} - \bar{\beta}_1 - \overline{\beta_2 x}) + \beta_{2j}x_{ij}) + (\varepsilon_{ij} - \bar{\varepsilon}) \\ &= \beta^*_{1j} + \beta_{2j}x_{ij}) + \varepsilon^*_{ij}, \end{aligned}$$

where $\beta^*_{1j} = \beta_{1j} - \bar{\beta}_1 - \overline{\beta_2 x}$, $\varepsilon^*_{ij} = \varepsilon_{ij} - \bar{\varepsilon}$, and

$$\overline{\beta_2 x} = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{N_j}\beta_{2j}x_{ij}.$$

The equation for the intercept in the transformed model becomes

$$\begin{aligned} \beta_{1j} &= \gamma_1 + \gamma_2 w_j + u_{1j} - \gamma_1 - \gamma_2\bar{w}_{(1)} - \bar{u}_1 - \gamma_3\bar{x} - \gamma_4\overline{wx} - \overline{u_2 x} \\ &= (-\gamma_2\bar{w}_{(1)} - \gamma_3\bar{x} - \gamma_4\overline{wx}) + \gamma_2 w_j + (u_{1j} - \bar{u}_1 - \overline{u_2 x}) \\ &= \gamma^*_1 + \gamma_2 w_j + u^*_{1j}, \end{aligned}$$

where

$$\begin{aligned} \gamma^*_1 &= -\gamma_2\bar{w}_{(1)} - \gamma_3\bar{x} - \gamma_4\overline{wx} \\ u^*_{1j} &= u_{1j} - \bar{u}_1 - \overline{u_2 x}, \end{aligned}$$

and $\overline{wx}$ and $\overline{u_2 x}$ are defined analogous to $\overline{\beta_2 x}$, and $\bar{u}_1$ is defined analogous to $\bar{w}_{(1)}$. If the fixed parameters are unrestricted, the transformation leaves the fixed part unchanged. We are not certain that the estimation results when treating ε^*_{ij} and u^*_{1j} as i.i.d. random variables with zero means are equivalent to the estimation results using the uncentered outcome variable, with the appropriate transformation of the fixed intercept γ_1 into γ^*_1. In the OLS regression case the results are equivalent, as stated above, but in this model they may not be. However, even if they are different, the differences will typically be very small, because ε^*_{ij} and u^*_{1j} still have mean zero, and thus their sample averages will also be close to zero, so that the differences between ε^*_{ij} and ε_{ij}, and the differences between u^*_{1j} and u_{1j} will be negligible, especially in large samples. Nevertheless, given these complications, it seems easier not to center the outcome variable. The interpretation of the model parameters will usually not be more difficult for an uncentered outcome variable.

2.3.3 Level-2 centering

Level-2 centering means centering within contexts, i.e., subtracting the group mean. Sometimes there is a theoretical rationale for assuming that the outcome variable does not depend on the value of an explanatory variable by itself, but on the relative value of this variable within the group. Some examples of situations in which this may be the case are given in Kreft et al. (1995, pp. 17–20) and Snijders and Bosker (1999, p. 81). Here we will first study the technical consequences of Level-2 centering and then make some remarks about substantive issues of model specification and their consequences.

Let us again consider (2.13) and see whether we can transform the model such that the Level-2-centered variable $x^{\bullet}_{ij} = x_{ij} - \bar{x}_j$ becomes the explanatory variable, where $\bar{x}_j$ denotes the sample mean of x within the j-th Level-2 unit. Thus, we write

$$\begin{aligned} y_{ij} &= \beta_{1j} + \beta_{2j}(x_{ij} - \bar{x}_j) + \beta_{2j}\bar{x}_j + \varepsilon_{ij} \\ &= \beta_{1j} + \beta_{2j}x^{\bullet}_{ij} + \beta_{2j}\bar{x}_j + \varepsilon_{ij} \\ &= \beta_{1j} + \beta_{2j}x^{\bullet}_{ij} + \beta_{3j}\bar{x}_j + \varepsilon_{ij}, \end{aligned} \tag{2.14}$$

with β_{3j} a (possibly) random slope. If it is assumed that (2.13) is the “true” model, then $\beta_{3j} = \beta_{2j}$, whereas replacing x_{ij} by its Level-2-centered counterpart in (2.13) implicitly imposes the assumption $\beta_{3j} = 0$. The term $\beta_{3j}\bar{x}_j$ is a random term that varies over Level-2 units, but is constant within each Level-2 unit. Therefore, as with grand mean centering, we may try to include it in the random intercept. This gives

$$\begin{aligned} y_{ij} &= \beta^{\bullet}_{1j} + \beta_{2j}x^{\bullet}_{ij} + \varepsilon_{ij} \\ \beta^{\bullet}_{1j} &= \beta_{1j} + \beta_{3j}\bar{x}_j \\ &= \gamma_1 + \gamma_2 w_j + u_{1j} + \gamma_3\bar{x}_j + \gamma_4 w_j\bar{x}_j + u_{2j}\bar{x}_j. \end{aligned}$$

The penultimate term introduces a Level-2 interaction variable, say, $v_j = w_j\bar{x}_j$, which was not present in the original model. If the original and transformed model are to be equivalent, then this means that v_j should also have been included in the original model. Hence, this would seem to give the advice that the intercept equation should always include the interactions (i.e., products) of all Level-2 variables with all group averages of Level-1 variables, just like we argued above that all Level-2 explanatory variables shoud be included in the intercept equation. However, in the current situation there may be strong substantive reasons why including these products would not be included. Moreover, inclusion of these products still does not render both models equivalent, because the term $u_{2j}\bar{x}_j$ introduces a Level-2 heteroskedastic error term, which does not fit into the MLA model specification. Consequently, viewing $\beta_{3j}\bar{x}_j$ as part of the random intercept does not make the uncentered and Level-2-centered model specifications equivalent.

To obtain a model specification that is equivalent with uncentered and Level-2-centered explanatory variables, we must take (2.14) as the basis, with

$$\beta_{3j} = \gamma_5 + \gamma_6 w_j + u_{3j}.$$

That is, we must include all within-group averages of the Level-1 explanatory variables as additional explanatory variables at Level-1. If the parameters are not restricted, then the resulting model satisfies the requirement that the models with and without Level-2 centering are equivalent. However, in many cases it is substantively highly questionable whether inclusion of the within-group averages as additional explanatory variables makes any sense. If they are excluded, then, as shown above, Level-2 centering leads to a different special case (different implicit restrictions on β_{3j}) than not centering, and the models are different, and centering is not innocuous.

As with grand mean centering, we can also study Level-2 centering of the w_j and y_{ij}. Clearly, Level-2 centering of a Level-2 variable like w_j is not useful, because this transformed variable then becomes zero and drops out of the model. Level-2 centering of the outcome variable gives

$$\begin{aligned}
y^{\bullet}_{ij} &= y_{ij} - \bar{y}_j \\
&= \beta_{1j} + \beta_{2j}x_{ij} + \varepsilon_{ij} - \beta_{1j} - \beta_{2j}\bar{x}_j - \bar{\varepsilon}_j \\
&= \beta_{2j}(x_{ij} - \bar{x}_j) + (\varepsilon_{ij} - \bar{\varepsilon}_j) \\
&= \beta_{2j}x^{\bullet}_{ij} + \varepsilon^{\bullet}_{ij},
\end{aligned}$$

with obvious notation. If x_{ij} is replaced by $x^{\bullet}_{ij}$ in the second of these equations, then the result is unaltered. Therefore, Level-2 centering of the outcome variable leads to a transformed model that is unaffected by Level-2 centering of the Level-1 explanatory variable. If one is uncertain whether Level-2 centering of the explanatory variables should be done or not, then Level-2 centering of the outcome variable solves the problem. In stating this, we have ignored the difference between ε_{ij} and $\varepsilon^{\bullet}_{ij}$, as usual.

The price of Level-2 centering of the outcome variable is that the intercept β_{1j} has dropped out of the equation. In many social science applications, the intercept is not very interesting, but in a multilevel model, the Level-1 intercept is usually important, if only because the main effects of the Level-2 explanatory variables are channeled through the random intercept. Therefore, in many cases one would not be willing to reduce the model by Level-2 centering of the outcome variable. Then, given the non-equivalence of the resulting models, the decision whether or not to use Level-2 centering must (primarily) be based on substantive arguments.

There are some additional issues with Level-2 centering that have not been stressed in the cited literature. Even if there is a good substantive reason to assume that the group-mean centered variable is the most relevant explanatory variable, one would assume that the really relevant variable would not be x_{ij} with the sample average subtracted, but with the population mean $\bar{X}_j$, say, subtracted. Except for some kinds of laboratory experiments, it does not make sense to assume that the outcome only depends on the average of x for the units in the sample and not on the units not included in the sample. Thus, the correct model would be

$$\begin{aligned}
y_{ij} &= \beta_{1j} + \beta_{2j}(x_{ij} - \bar{X}_j) + \varepsilon_{ij} \\
&= \beta_{1j} + \beta_{2j}(x_{ij} - \bar{x}_j) + \beta_{2j}(\bar{x}_j - \bar{X}_j) + \varepsilon_{ij}.
\end{aligned}$$

If we would consider the x-es to be random variables with some distribution, $\bar{x}_j$ would converge to $\bar{X}_j$ with an increasing sample size within Level-2 unit j, and with moderately large within-groups sample sizes, we would not expect the term $\bar{x}_j - \bar{X}_j$ to have an important effect on the results. However, it leads to a couple of interesting questions from a theoretical perspective. The difference between $\bar{x}_j$ and $\bar{X}_j$ leads to a measurement error problem for which it seems to matter whether the (sub)sample is assumed to be drawn from a finite (sub)population or from an infinite population.

However, the statistical theory of the multilevel model assumes that the x-es are given constants. This means that they are either chosen by an experimental researcher (as in agricultural experiments), or they are essentially random but the analysis is done conditional on their outcomes. The latter is common in the social sciences. See De Leeuw and Kreft (1995, p. 173) for a discussion of this issue. The problem with conditioning on the x-es is that $\bar{x}_j - \bar{X}_j$ is a constant which is unknown, typically nonzero, and different for different Level-2 units. Except for some special cases, this would normally imply that according to the standard multilevel theory, the estimators are inconsistent. However, this is also not the case here, because, although $\bar{x}_j - \bar{X}_j$ is considered a constant over repeated sampling with the same sample size, it varies with increasing sample size, converging to zero at a rate of $1/\sqrt{N_j}$. So the estimators seem to be consistent after all, but we are unaware of the fine details of its proof.

A more important issue associated with Level-2 centering is that it may well be the case that instead of $x_{ij} - \text{mean}(x)_j$, the relevant variable is another relative measure, such as $x_{ij} - \text{median}(x)_j$, $x_{ij}/\text{mean}(x)_j$, $x_{ij}/\text{median}(x)_j$, $[x_{ij} - \text{mean}(x)_j]/\text{sd}(x)_j$, and so forth. Most theory that suggests that a Level-2-centered variable should be used is not strong enough to rule out these alternative specifications. Because it matters which one is taken, one should ideally either come up with a good reason why a specific form should be chosen, or perform a detailed specification analysis. However, `MLA` currently only supports centering with respect to the mean, so other transformations must already have been made in the data file that is read by `MLA`.

2.4 Ordinary least squares estimators

2.4.1 Within-group models

In this section, we will use the notation (2.1)–(2.2). Based on (2.1), ordinary least squares estimates for β_j are given by

$$\widehat{\beta}_j = (X_j' X_j)^{-1} X_j' y_j, \tag{2.15}$$

and the estimated standard errors of the elements of $\widehat{\beta}_j$ are the square roots of the diagonal elements of the covariance matrix given by

$$\widehat{\text{cov}}(\widehat{\beta}_j) = \widehat{\sigma}_j^2 (X_j' X_j)^{-1}, \tag{2.16}$$

where $\widehat{\sigma}_j^2 = \frac{1}{N_j - q} \widehat{\varepsilon}_j' \widehat{\varepsilon}_j$, $\widehat{\varepsilon}_j = y_j - X_j \widehat{\beta}_j$), and q is the dimension of β_j. Because these are standard regression estimators, the usual regression theory applies within each Level-2 unit, given that β_j is now regarded as a parameter vector instead of as a random vector. In particular, this means that under the usual assumptions, $\widehat{\beta}_j$ is unbiased with covariance matrix $\sigma_j^2 (X_j' X_j)^{-1}$, where σ_j^2 is consistently (and unbiasedly) estimated by $\widehat{\sigma}_j^2$. These results do not require normality and, as is clear from this discussion, the variances of the Level-1 residuals are allowed to be different in different groups.

In typical multilevel datasets, the sample size within a given Level-2 unit may be quite small compared to ordinary regression analyses. Thus, although unbiased and consistent, these OLS estimators may not be very precise and the estimated covariance matrices may also not be very precise.

Although it is not very common in linear regression, we can also compute standard errors for $\widehat{\sigma}_j^2$. Unlike the covariance matrix of $\widehat{\beta}_j$, the variance of $\widehat{\sigma}_j^2$ also depends on whether the normality assumption holds. The standard error that is printed in `MLA` is

$$\text{se}(\widehat{\sigma}_j^2) = \frac{2\widehat{\sigma}_j^4}{N_j - q},$$

which is only correct when the normality assumption holds. Future versions of `MLA` will also compute robust standard errors, which are correct under nonnormality.

2.4.2 One-step `OLS` (total model)

From Equation (2.6) the term $Zu + \varepsilon$ can be considered the random part of the equation. Taking the total residuals

$$r = Zu + \varepsilon, \tag{2.17}$$

leaves, after substitution,

$$y = X\gamma + r. \tag{2.18}$$

Now, γ can be estimated consistently using ordinary least squares. Notice that grouping is ignored. Estimates for γ are given by

$$\widehat{\gamma} = (X'X)^{-1}X'y. \tag{2.19}$$

Using the estimated residuals $\widehat{r} = y - X\widehat{\gamma}$, the estimate of the variance of the elements of r can be obtained by

$$\widehat{\sigma}_r^2 = \frac{1}{N-p}\sum_{i=1}^{N}\widehat{r}_i^2, \tag{2.20}$$

where p is the dimension of γ. This estimate $\widehat{\sigma}_r^2$ is the one-step `OLS` estimate of the variance of the residuals. The usual standard errors for $\widehat{\gamma}$ and $\widehat{\sigma}_r^2$ are, respectively,

$$\widehat{\mathrm{se}}(\widehat{\gamma}_l) = \sqrt{[\widehat{\sigma}_r^2(X'X)^{-1}]_{ll}}, \tag{2.21}$$

$$\widehat{\mathrm{se}}(\widehat{\sigma}_r^2) = \widehat{\sigma}_r^2\sqrt{\frac{2}{N-p}}. \tag{2.22}$$

These are, however, only correct if the classical assumptions of i.i.d. normally distributed residuals are correct, which is generally not the case for multilevel data. Again, future versions of `MLA` will provide more robust standard errors.

2.4.3 Two-step `OLS` (total model)

With the two-step `OLS`, the same estimates $\widehat{\gamma}$ are used as with the one-step `OLS`, see (2.19). The total residuals for every group j can be divided into a Level-2 and a Level-1 part. This was already done in Equation (2.17). Using ordinary least squares, estimates for the Level-2 random components, u, can be obtained by

$$\widehat{u}_j = (Z_j'Z_j)^{-1}Z_j'\widehat{r}_j. \tag{2.23}$$

The estimate for the covariance matrix Θ of u becomes

$$\widehat{\Theta} = \frac{1}{J}\sum_{j=1}^{J}\widehat{u}_j\widehat{u}_j'. \tag{2.24}$$

Under normality, the estimated covariances of the elements of $\widehat{\Theta}$ can be obtained by (Anderson, 1958, p. 161)

$$\widehat{\mathrm{cov}}(\widehat{\Theta}_{kl}, \widehat{\Theta}_{mn}) = (\widehat{\Theta}_{km}\widehat{\Theta}_{ln} + \widehat{\Theta}_{kn}\widehat{\Theta}_{lm})/J. \tag{2.25}$$

Consequently, the estimated standard errors of the elements of $\widehat{\Theta}$ are given by

$$\widehat{\mathrm{se}}(\widehat{\Theta}_{kl}) = \sqrt{(\widehat{\Theta}_{kk}\widehat{\Theta}_{ll} + \widehat{\Theta}_{kl}^2)/J}. \tag{2.26}$$

By first computing the residuals $\widehat{\varepsilon}$,

$$\widehat{\varepsilon} = \widehat{r} - Z\widehat{u}, \tag{2.27}$$

the estimate for σ_ε^2 becomes

$$\widehat{\sigma}_\varepsilon^2 = \frac{1}{N}\sum_{i=1}^{N}\widehat{\varepsilon}_i^2. \tag{2.28}$$

This estimate $\widehat{\sigma}_\varepsilon^2$ is the two-step OLS estimate of the variance of the elements of ε. The estimated standard error for $\widehat{\sigma}_\varepsilon^2$ becomes, analogous to Equation (2.22),

$$\widehat{\mathrm{se}}(\widehat{\sigma}_\varepsilon^2) = \widehat{\sigma}_\varepsilon^2\sqrt{\frac{2}{N-q}}. \tag{2.29}$$

As with previously presented standard errors, the standard errors given in this section will be supplemented with robust ones in future versions of MLA.

All the estimators in this section are consistent if $J \longrightarrow \infty$ and $N_j \longrightarrow \infty$ for each j. Although this may be unrealistic, these estimators may be good initial estimators (starting values) for maximum likelihood estimators. In some cases, the differences between these estimators and the maximum likelihood estimators is small, and therefore, these estimators can be used as well (Kreft, 1996; Van der Leeden & Busing, 1994).

2.5 Full Information Maximum Likelihood (FIML)

One of the most important parts of the program consists of the maximum likelihood estimation. This estimation method was chosen for its desirable properties, such as consistency and efficiency. In maximum likelihood estimation, given the observations, parameters are found that maximize the likelihood function (Mood et al., 1974). This is the same as minimizing the minus-log-likelihood function. Assuming normally distributed errors, the density of y_j, given X_j and Z_j, is

$$f(y_j|X_j, Z_j) = \frac{1}{(2\pi)^{N_j/2}(\det V_j)^{1/2}}e^{-\frac{1}{2}(y_j - X_j\gamma)'V_j^{-1}(y_j - X_j\gamma)},$$

so that the contribution of Level-2 unit j to the minus-log-likelihood function is

$$\begin{aligned} L_j &= -\log f(y_j|X_j, Z_j) \\ &= \frac{N_j}{2}\log(2\pi) + \frac{1}{2}\log\det V_j + \frac{1}{2}(y_j - X_j\gamma)'V_j^{-1}(y_j - X_j\gamma) \end{aligned}$$

and the minus-log-likelihood function for the whole sample is simply the sum of all Level-2 units j, $L = \sum_{j=1}^{J} L_j$. This is the function that has to be minimized with respect to the parameters to obtain maximum likelihood estimators. Specifically, it will produce a set of fixed parameter estimates, $\widehat{\gamma}$, and a set of random parameter estimates, $\widehat{\Theta}$ for the second level and $\widehat{\sigma}_\varepsilon^2$ for the first level. Details can be found in Appendix A.

The asymptotic covariance matrix of the estimators is derived from the matrix of second derivatives of L (the Hessian matrix). This covariance matrix is used for the standard errors for both fixed and random parameters.

For a detailed description of the derivations used and an extensive discussion of the computational formulas used in the program, Appendix A contains all information.

2.6 Restricted Maximum Likelihood (REML)

Restricted maximum likelihood (REML) estimators are the maximum likelihood estimators of a transformed version of (2.6). In particular, let K be an $N \times (N - p)$ matrix of rank $N - p$ such that $K'X = 0$. Then

$$w = K'y = K'Zu + K'\varepsilon$$

does not depend on γ. Its distribution is given by

$$w \sim \mathcal{N}(0, K'VK)$$

and its $(-2\log)$ likelihood equation is

$$L_w = \log\det(K'VK) + w'(K'VK)^{-1}w.$$

Note that L_w does not depend on the fixed parameters γ. The REML estimators of the random parameters Θ and σ_ε^2 are obtained by minimizing L_w instead of the full $(-2\log)$ likelihood L as given in section 2.5. It can be shown that the resulting estimators do not depend on the specific choice of K, any $N \times (N - p)$ matrix K of rank $N - p$ such that $K'X = 0$ will lead to the same estimators.

It turns out that REML estimators are typically less biased than FIML estimators. This is one of the reasons why REML is often preferred. Asymptotically, however, the difference between FIML and REML becomes negligible.

Because L_w does not depend on γ, REML estimators of γ do not exist. Usually, however, it is desirable to estimate γ as well. A good estimator of γ can be obtained as follows. In (2.18), the two-level model was written as a linear regression model in which the covariance matrix V of the residuals r is not a scalar multiple of the identity matrix. If V were known, the best linear unbiased estimator of γ would be the generalized least squares (GLS) estimator

$$\widehat{\gamma}_{\text{GLS}} = (X'V^{-1}X)^{-1}X'V^{-1}y,$$

which is also the maximum likelihood estimator if normality is assumed. Because V is unknown, this estimator cannot be computed. However, we can estimate V consistently by plugging in the REML estimators of Θ and σ_ε^2, giving the estimator $\widehat{V}_{\text{REML}}$. The resulting *feasible generalized least squares* (FGLS) estimator is

$$\widehat{\gamma}_{\text{FGLS}} = (X'\widehat{V}_{\text{REML}}^{-1}X)^{-1}X'\widehat{V}_{\text{REML}}^{-1}y.$$

This estimator can also be called *two-step maximum likelihood* (Murphy & Topel, 1985), because in the first step, Θ and σ_ε^2 are estimated by minimizing the restricted $(-2\log)$ likelihood L_w and in the second step, the likelihood of y is maximized with respect to γ, conditional on the first-step estimators of Θ and σ_ε^2. In the terminology of Gong and Samaniego (1981) and Parke (1986), this is called *pseudo maximum likelihood*, but this term is ambiguous, because other authors use it for maximizing an "incorrect" likelihood, such as maximizing L when the normality assumption is not met. This is called *quasi maximum likelihood* by White (1982). Note that in the output of MLA, the two-step maximum likelihood estimates of γ are somewhat incorrectly printed under the "restricted maximum likelihood" heading.

2.7 Residuals

The total residuals are given by Equation (2.18),

$$\widehat{r} = y - X\widehat{\gamma}.$$

In this equation, estimated residuals $\widehat{r}$ are based on the fixed parameter estimates $\widehat{\gamma}$ from the maximum likelihood estimation, although other estimates of γ could be used as well.

The raw residuals for the first level are taken from the within-group model (2.1),

$$\widehat{\varepsilon}_j = y_j - X_j\widehat{\beta}_j, \tag{2.30}$$

where the estimates $\widehat{\beta}_j$ are the `OLS` estimates from (2.15). Using the between-group model from Equation (2.2) and the `OLS` estimates from Equation (2.15), the Level-2 raw residuals are

$$\widehat{u}_j = \widehat{\beta}_j - W_j\widehat{\gamma}, \tag{2.31}$$

where $\widehat{\gamma}$ stems from Equation (2.19).

The Level-2 shrunken residuals are given by

$$\widehat{u}_j = (Z_j\widehat{\Theta})'[Z_j\widehat{\Theta}Z_j' + \widehat{\sigma}^2_\varepsilon I_{N_j}]^{-1}\widehat{r}_j. \tag{2.32}$$

where $\widehat{r}_j$ contains the total full information maximum likelihood residuals for group j (i. e., $\widehat{r}_j = y_j - X_j\widehat{\gamma}$, where $\widehat{\gamma}$ is the `FIML` estimator of γ), and $\widehat{\Theta}$ and $\widehat{\sigma}^2_\varepsilon$ are the `FIML` estimators of Θ and σ^2_ε, respectively. The formula is computationally rather inefficient. Therefore, the following more efficient formulas will be used. One can write

$$(Z_j\widehat{\Theta})' = \widehat{\Theta}Z_j'$$

and taking

$$\begin{aligned} \widehat{V}_j &= Z_j\widehat{\Theta}Z_j + \widehat{\sigma}^2_\varepsilon I_{N_j} && \text{(from (2.5))},\\ Z_j'\widehat{V}_j^{-1} &= \widehat{\sigma}_\varepsilon^{-2}\widehat{G}_j^{-1}Z_j' && \text{(from (A.12))}, \end{aligned}$$

where

$$\widehat{G}_j = I_q + Z_j'Z_j\widehat{\Theta}/\widehat{\sigma}^2_\varepsilon, \tag{2.33}$$

then

$$\begin{aligned} \widehat{u}_j &= \widehat{\Theta}Z_j'\widehat{V}_j^{-1}(y_j - X_j\widehat{\gamma}_j)\\ &= \widehat{\sigma}_\varepsilon^{-2}\widehat{\Theta}\widehat{G}_j^{-1}(Z_j'y_j - Z_j'X_j\widehat{\gamma}_j) \end{aligned}$$

Finally, the shrunken residuals for Level-1 follow from (2.17),

$$\widehat{\varepsilon} = \widehat{r} - Z\widehat{u}. \tag{2.34}$$

2.8 Posterior means

The posterior means are the shrunken estimators of β_j. They are the expected values of the β_j, given the data and the maximum likelihood estimates of γ, Θ, and σ^2_ε. They are derived from the shrunken residuals and their formula is

$$\widehat{\beta}_j = W_j\widehat{\gamma} + \widehat{u}_j, \tag{2.35}$$

where $\widehat{\gamma}$ is the estimate obtained by full information maximum likelihood and $\widehat{u}_j$ is taken from (2.32). This can easily be shown to be equal to

$$\widehat{\beta}_j = (I_q - \widehat{\Lambda}_j)(W_j\widehat{\gamma}) + \widehat{\Lambda}_j\widehat{\beta}_j^{\texttt{OLS}}, \tag{2.36}$$

where $\widehat{\beta}_j^{\texttt{OLS}}$ is the within-group `OLS` estimator (2.15) of β_j, $\widehat{\Lambda}_j = \widehat{\Theta}X_j'\widehat{V}_j^{-1}X_j$, $\widehat{V}_j = \widehat{\sigma}_\varepsilon^2 I_{N_j} + X_j\widehat{\Theta}X_j'$, and the notation is of (2.1) and (2.2). Thus, from (2.36), $\widehat{\beta}_j$ can be seen as a matrix-weighted average of the within-group estimator $\widehat{\beta}_j^{\texttt{OLS}}$ and the estimated prior expectation $W_j\widehat{\gamma}$ of β_j, the former being unbiased and the latter being more efficiently estimated. The more efficient $\widehat{\beta}_j^{\texttt{OLS}}$ is estimated, the more (matrix) weight it gets, and the closer the posterior means are to the within-group estimates. $\widehat{\Lambda}_j$ can be called an estimated "reliability" matrix (cf. Bryk & Raudenbush, 1992, p. 43).

2.9 Diagnostics

Currently, the only option for diagnostics performed by `MLA` apart from the descriptive statistics, is outlier detection. Although the term outlier seems to be unambiguous, this is not completely true. An outlier is considered to be a deviant observation in the data and not a deviant residual after model estimation. However, a procedure fitting outliers in the data as residual outliers is considered to be a robust procedure. Outliers are in `MLA` detected using residuals. So, we expect `MLA` to be robust against data outliers and therefore look for residual outliers. More research in the field of robustness for multilevel models would be useful, however.

The detection of outliers differs for Level-1 and Level-2 outliers. For both levels, the shrunken residuals are considered. For the first level, the quotients

$$\frac{\widehat{\varepsilon}_{ij}}{\sqrt{\widehat{\sigma}^2(\widehat{\varepsilon})}} \tag{2.37}$$

are calculated, where

$$\widehat{\sigma}^2(\widehat{\varepsilon}) = \frac{1}{N}\sum_{i,j}\widehat{\varepsilon}_{ij}^2$$

is the variance of the Level-1 residuals. Residuals will be displayed whenever the quotient (2.37), when compared to a standard normal distribution, has a p-value less than some (possibly) user-specified value. The default value is 0.1.

For the Level-2 outliers, the Mahalanobis distances of the Level-2 residuals to their theoretical mean of zero are calculated by

$$M_j = u_j'\widehat{\Theta}^{-1}u_j.$$

Now, residuals are displayed for which M_j is larger than the critical value corresponding to a (possibly) user-specified p-value of a chi-square distribution with q degrees of freedom, where q is the dimension of u. This p-value is the same as for the Level-1 outliers.

2.10 Simulation

The maximum likelihood theory discussed so far is based on a few assumptions, the most important of which are:

- The model (i. e., the conditional expectation $X\gamma$ and covariance matrix V) is correctly specified. The standard errors, t-values, exceedance probabilities, and likelihood ratio tests were derived under the condition that at least the (most general) model that is being estimated is correct in the population.

- The Level-1 (ε) and Level-2 (u) random errors are normally distributed. The likelihood function was derived under this assumption, and therefore, the FIML estimators and the estimators of their standard errors depend on it.

- The sample size is large. More specifically, the properties of the maximum likelihood estimators, such as their consistency, their (asymptotic) efficiency, and their (asymptotic) normal distribution, as well as the formulas for their standard errors were derived under the assumption that the sample size goes to infinity ($N \longrightarrow \infty$).

In practice, these assumptions will not be completely satisfied. One can only hope that they are met approximately. To be able to get an indication of how severe the finite sample size and possible nonnormality influence the results, the MLA program offers simulation options. In this section, the theory underlying these simulation options will be described. This focus will be on the possible bias of the estimates and on the possibly incorrect standard errors. More subtle information can, however, be extracted from the program by using a file to write the simulation results to.

The *bias* of an estimator $\widehat{\theta}$ of some parameter θ is defined as the difference between the expected value of the estimator and the true value of the parameter. A desirable property of an estimator is *unbiasedness*, which means that its bias is zero. In the maximum likelihood theory discussed so far, however, it was only stated that the FIML estimators are *consistent*. This means that as the sample size gets larger, the mean of the estimator converges to the true parameter value and its variance decreases to zero. Informally speaking, the estimator comes closer to the true parameter value as sample size gets larger. This is, of course, a highly desirable property, but it does not ensure that the estimator is unbiased in finite samples. In fact, maximum likelihood estimators are in many models and situations biased in finite samples. For a general class of regression models including multilevel models, however, Magnus (1978) proved that the maximum likelihood estimators of the *fixed* regression coefficients are unbiased. On the other hand, Busing (1993) showed in a Monte Carlo simulation study that the maximum likelihood estimators of the *random* parameters in multilevel models are biased.

The standard errors of the maximum likelihood estimators that are reported by MLA are derived from asymptotic theory. This means that they are based on the idea that as the sample size goes to infinity, the distribution of the estimators will converge to a (multivariate) normal distribution with a certain covariance matrix (see Appendix A). The reported standard errors that are the square roots of the diagonal elements of this matrix. The exceedance probabilities of the according t-values that are reported are based on the approximation of the distribution of the estimators by the normal distribution. In finite samples, this approximation may not be very good. The true standard errors may be quite different from the reported ones based on asymptotic theory, and the distributions of the estimators may not be normal. In fact, Busing (1993) showed in his simulation study that the distributions of the random parameters can be severely skewed. As mentioned above, however, the focus is on the bias and the standard errors and not on the specific distribution.

2.10.1 The jackknife

The jackknife was originally introduced by Quenouille (1949, 1956) to estimate the bias of an estimator and to correct for it. Tukey (1958) proposed an accompanying estimator for the variance of the estimator, and hence for its standard error.

The idea of the jackknife is as follows. Consider an independently and identically distributed sample of size N from some distribution and an estimator $\widehat{\theta}_N$ of a parameter θ obtained from this sample. Furthermore, consider removing a group of m observations from the sample, and let $\widehat{\theta}_{N-m}$ be the estimator of the same parameter θ based on this sample of size $N - m$. The difference between $\widehat{\theta}_N$ and $\widehat{\theta}_{N-m}$ can then be used to estimate the bias of $\widehat{\theta}_N$ and this estimate can be used to obtain the bias-corrected jackknife estimator $\widehat{\theta}_J$. It is known that the bias of $\widehat{\theta}_J$ is generally of order N^{-2} if m is relatively small compared to N. This is typically much smaller than the bias of $\widehat{\theta}_N$, which is generally of order N^{-1}.

Obviously, there are many possibilities for selecting a group of observations of size m from the sample. If m is equal for each group, the simplest case is obtained for $m = 1$. Now, the sample is divided into N "groups" of size one, i.e., the N observations. In all other cases with $m > 1$ and N a multiple of m, the sample is divided into g mutually exclusive groups of size m, with $g = N/m$. We will first give the details of the standard jackknife procedures for these situations and then describe a grouped jackknife procedure that is more suitable for typical multilevel data. Justifications can be found in the standard jackknife literature (e.g., Shao & Tu, 1995) or as special cases of the discussion in section 2.10.1 below.

Delete-1 jackknife

Suppose $\widehat{\theta}_N$ is an estimator of θ based on a sample of size N. Now, remove the i-th observation from the sample, and let $\widehat{\theta}_{(i)}$ be the estimator of θ based on a sample size of $N - 1$. The delete-1 jackknife estimator of θ is now given by

$$\widehat{\theta}_{J(1)} = N\widehat{\theta}_N - (N-1)\bar{\theta}_{(1)}, \tag{2.38}$$

where $\bar{\theta}_{(1)} = N^{-1}\sum_{i=1}^{N}\widehat{\theta}_{(i)}$.

The delete-1 jackknife variance estimator (Tukey, 1958), based on the *pseudo values*

$$\tilde{\theta}_{(i)} = N\widehat{\theta}_N - (N-1)\widehat{\theta}_{(i)}, \qquad i = 1, \dots, N,$$

is given by

$$\begin{aligned} \widehat{\sigma}^2_{J(1)} &= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N-1}\left(\tilde{\theta}_{(i)} - \frac{1}{N}\sum_{k=1}^{N}\tilde{\theta}_{(k)}\right)^2, \\ &= \frac{N-1}{N}\sum_{i=1}^{N}\left(\widehat{\theta}_{(i)} - \bar{\theta}_{(1)}\right)^2. \end{aligned} \tag{2.39}$$

As mentioned above and discussed in more detail in section 2.10.1, the bias of $\widehat{\theta}_{J(1)}$ is typically $O(N^{-2})$, whereas the bias of $\widehat{\theta}_N$ is typically $O(N^{-1})$. Furthermore, $\widehat{\sigma}^2_{J(1)}$ is a consistent estimator of the asymptotic variance of both $\widehat{\theta}_N$ and $\widehat{\theta}_{J(1)}$.

Delete-*m* jackknife

Suppose the sample is divided into g mutually exclusive and independent groups of (equal) size m ($m > 1$), where $m = N/g$. Now remove the m observations of group j from the sample, and let $\widehat{\theta}_{(j)}$ be the estimator of θ based on the corresponding reduced sample of size $N - m$. The delete-*m* jackknife (or grouped jackknife) estimator of θ is now given by

$$\widehat{\theta}_{J(m)} = g\widehat{\theta}_N - (g-1)\bar{\theta}_{(m)}, \tag{2.40}$$

with $\bar{\theta}_{(m)} = g^{-1}\sum_{j=1}^{g}\widehat{\theta}_{(j)}$. Hence, $\widehat{\theta}_{J(m)}$ is based on g estimators $\widehat{\theta}_{(j)}$ of θ, each based on a subsample of size $N - m$. Clearly, for $m = 1$, (2.40) reduces to (2.38).

The delete-*m* jackknife variance estimator is defined similarly to (2.39). It is based on the pseudo values

$$\tilde{\theta}_{(j)} = g\widehat{\theta}_N - (g-1)\widehat{\theta}_{(j)}, \qquad j = 1, \ldots, g,$$

and given by

$$\begin{aligned}\widehat{\sigma}^2_{J(m)} &= \frac{1}{g}\sum_{j=1}^{g}\frac{1}{g-1}\left(\tilde{\theta}_{(j)} - \frac{1}{g}\sum_{k=1}^{g}\tilde{\theta}_{(k)}\right)^2, \\ &= \frac{g-1}{g}\sum_{j=1}^{g}\left(\widehat{\theta}_{(j)} - \bar{\theta}_{(m)}\right)^2.\end{aligned} \tag{2.41}$$

The mathematics leading to (2.38)–(2.41) can be found in the standard jackknife literature. For example, Shao and Tu (1995) provide a systematic introduction to the theory of the jackknife, including a discussion of its theoretical properties.

Delete-m_j jackknife

Above, we discussed the classic jackknife approach to estimating the bias of an estimator and to obtain a bias-corrected version of this estimator. Using the pseudo values, an accompanying estimator for the variance of the (original or bias-corrected) estimator, and hence for the standard error, can be obtained as well.

The jacknife version we discussed is based on subsamples obtained from the original sample by successively removing mutually exclusive groups of observations of size m. Furthermore, it relies on the assumption of independently and identically distributed observations. Both features influence the formulation of a jackknife resampling scheme for multilevel data and models.

The independence assumption restricts the application of the jackknife to the highest level in the data. In the two-level case, independence can only be assumed for the groups. Within the groups, data are dependent. Consequently, a multilevel jackknife approach must be based on subsamples obtained by removing complete Level-2 units. In fact, Wolter (1985, section 4.6) already stated that the delete-*m* jackknife can be used in cluster sampling, when the data within clusters are dependent. In multilevel data, however, groups are usually not of *equal* size m. Therefore, to make the jackknife suitable for multilevel data and models, the delete-*m* jackknife needs to be generalized to a grouped jackknife for unequal group sizes, called the delete-m_j jackknife by Busing, Meijer, and Van der Leeden (1999).

To apply the delete-*m* jackknife (with $m > 1$, and N a multiple of m), the sample is divided into g mutually exclusive groups of size m, with $g = N/m$. In multilevel analysis, the sample is divided into J groups of (usually) varying size N_j, that is, N_j is not equal for each group and

N/N_j will not necessarily be equal to J. As a result, the formulas discussed earlier have to be adapted slightly. Let $\widehat{\theta}_{(j^*)}$ be an estimator of θ based on a sample from which group j with size N_j is removed. The delete-m_j jackknife estimator of θ is now given by

$$\widehat{\theta}_{J(m_j)} = J\widehat{\theta}_N - \sum_{j=1}^{J}\left(1 - \frac{N_j}{N}\right)\widehat{\theta}_{(j^*)}. \tag{2.42}$$

The estimator $\widehat{\theta}_{J(m_j)}$ can be justified as follows. Consider an estimator $\widehat{\theta}_N$ of a parameter θ obtained from a sample of size N from some distribution. In general, the expected value of such estimators can be written as the true value θ_0 plus a power series expansion in $1/N$, that is,

$$E(\widehat{\theta}_N) = \theta_0 + \frac{b_1}{N} + \frac{b_2}{N^2} + \frac{b_3}{N^3} + \dots, \tag{2.43}$$

where $b_1, b_2, \dots$ are unknown constants, independent of sample size, and frequently not equal to zero (see, e.g., Quenouille, 1956; Schucany, Gray, & Owen, 1971). If $b_1 \neq 0$, the bias in (2.43) is clearly of order N^{-1}. Let $h_j = N/N_j$. Then, the total sample size can be written as $N = N_j h_j$. Hence,

$$E(h_j\widehat{\theta}_N) = h_j\theta_0 + \frac{b_1}{N_j} + \frac{b_2}{h_j N_j^2} + \frac{b_3}{h_j^2 N_j^3} + \dots, \tag{2.44}$$

and

$$E(\widehat{\theta}_{(j^*)}) = \theta_0 + \frac{b_1}{(h_j - 1)N_j} + \frac{b_2}{(h_j - 1)^2 N_j^2} + \frac{b_3}{(h_j - 1)^3 N_j^3} + \dots. \tag{2.45}$$

Combining (2.44) and (2.45) gives

$$\begin{aligned} &E\left[h_j\widehat{\theta}_N - (h_j - 1)\widehat{\theta}_{(j^*)}\right] \\ &\quad = \theta_0 + \frac{b_2}{N_j^2}\left(\frac{1}{h_j} - \frac{1}{h_j - 1}\right) + \frac{b_3}{N_j^3}\left(\frac{1}{h_j^2} - \frac{1}{(h_j - 1)^2}\right) + \dots, \\ &\quad = \theta_0 - \frac{b_2}{N^2}\frac{h_j}{h_j - 1} - \frac{b_3}{N^3}\frac{h_j(2h_j - 1)}{(h_j - 1)^2} + \dots. \end{aligned} \tag{2.46}$$

Finally, to prevent loss of efficiency, the weighted average of the J possible estimators is used (cf. Quenouille, 1956). This gives

$$\begin{aligned} \widehat{\theta}_{J(m_j)} &= \sum_{j=1}^{J}\frac{N_j}{N}\left(h_j\widehat{\theta}_N - (h_j - 1)\widehat{\theta}_{(j^*)}\right), \\ &= J\widehat{\theta}_N - \sum_{j=1}^{J}\left(1 - \frac{N_j}{N}\right)\widehat{\theta}_{(j^*)}. \end{aligned} \tag{2.47}$$

The expectation of (2.47) is

$$\begin{aligned} &E\left[\sum_{j=1}^{J}\frac{N_j}{N}\left(h_j\widehat{\theta}_N - (h_j - 1)\widehat{\theta}_{(j^*)}\right)\right] \\ &\quad = \sum_{j=1}^{J}\frac{N_j}{N}\theta_0 - \sum_{j=1}^{J}\frac{1}{h_j}\left(\frac{b_2}{N^2}\frac{h_j}{h_j - 1}\right) - \sum_{j=1}^{J}\frac{1}{h_j}\left(\frac{b_3}{N^3}\frac{h_j(2h_j - 1)}{(h_j - 1)^2}\right) + \dots, \\ &\quad = \theta_0 - \frac{b_2}{N^2}\sum_{j=1}^{J}\frac{1}{h_j - 1} - \frac{b_3}{N^3}\sum_{j=1}^{J}\frac{2h_j - 1}{(h_j - 1)^2} + \dots, \end{aligned}$$

so that the bias is of order N^{-2} if $b_2 \neq 0$ and if N_j is relatively small compared to N.

The corresponding estimator of the variance of $\widehat{\theta}_{J(m_j)}$, based on the pseudo values

$$\tilde{\theta}_{(j^*)} = h_j \widehat{\theta}_N - (h_j - 1)\widehat{\theta}_{(j^*)}, \qquad j^* = 1, \ldots, J,$$

is given by

$$\begin{aligned}\widehat{\sigma}^2_{J(m_j)} &= \frac{1}{J} \sum_{j=1}^{J} \frac{1}{h_j - 1} \left(\tilde{\theta}_{(j^*)} - \widehat{\theta}_{J(m_j)} \right)^2 \\ &= \frac{1}{J} \sum_{j=1}^{J} \frac{1}{h_j - 1} \left(h_j \widehat{\theta}_N - (h_j - 1)\widehat{\theta}_{(j^*)} \right. \\ &\qquad\qquad \left. - J\widehat{\theta}_N + \sum_{k=1}^{J} \left(1 - \frac{N_k}{N} \right) \widehat{\theta}_{(k^*)} \right)^2 . \end{aligned} \tag{2.48}$$

Note that when all groups are of equal size, (2.40) follows from (2.47), that is, the delete-m_j jackknife estimator reduces to the delete-m jackknife estimator. Analogously, (2.48) reduces to the expression for the delete-m jackknife variance estimator (2.41).

2.10.2 Jackknife confidence intervals

The delete-m_j jackknife estimator and the delete-m_j jackknife variance estimator can be used to construct the jackknife normal confidence interval

$$\left[\widehat{\theta}_{J(m_j)} + z_{\frac{1}{2}\alpha} \widehat{\sigma}_{J(m_j)};\ \widehat{\theta}_{J(m_j)} + z_{1-\frac{1}{2}\alpha} \widehat{\sigma}_{J(m_j)} \right], \tag{2.49}$$

where $z_\alpha = \Phi^{-1}(\alpha)$ and $\Phi(\cdot)$ is the standard normal distribution function. The jackknife normal confidence interval relaxes the normality assumption for the data. However, the interval relies on the asymptotic normality of the estimators, which may in finite samples not be approximately satisfied (Busing, 1993). Other jackknife confidence intervals are not applicable or are probably worse, due to the limited use of the pseudo values.

2.10.3 The bootstrap

The *bootstrap* was introduced by Efron (1979) as an alternative to the jackknife. The idea of the bootstrap is that the empirical distribution function is a consistent estimator of the distribution function in the population. Let Z be a random variable with distribution function F, and let $\{z_1, z_2, \ldots, z_N\}$ be a random sample of size N from F. Now, the empirical distribution function $\widehat{F}_N$ in some point z is the proportion of z_i that are smaller than or equal to z:

$$\widehat{F}_N(z) = \frac{\#\{i : 1 \leq i \leq N | z_i \leq z\}}{N}. \tag{2.50}$$

If Z has a multivariate distribution, this formula has an obvious generalization and all subsequent formulas will also have obvious generalizations. It is known (e. g., Mood et al., 1974, p. 507) that, as $N \longrightarrow \infty$, $\widehat{F}_N(z) \longrightarrow F(z)$.

Let θ be a parameter associated with the distribution F, $\theta = \theta(F)$, and let $\widehat{\theta}$ be an estimator of θ from a sample, $\widehat{\theta} = \theta(z_1, z_2, \ldots, z_N) = \theta(\widehat{F}_N)$. The idea of the bootstrap is now to simulate the sampling and estimation process, where samples are drawn from $\widehat{F}_N$, which is completely known once the original sample is obtained. In the simulation, the distribution $\widehat{F}_N$ plays the role of F and $\widehat{\theta}$ plays the role of θ: Simulation samples $\{z_1^*, z_2^*, \ldots, z_N^*\}$ are drawn from $\widehat{F}_N$ and $\widehat{\theta}$ is estimated by θ^* in the same way θ was estimated by $\widehat{\theta}$.

Because $\widehat{F}_N \longrightarrow F$, it is assumed that the properties of the estimator θ^* based on the distribution $\widehat{F}_N$ give information about the properties of $\widehat{\theta}$ based on the distribution F. For example, the bias of θ^* based on the distribution $\widehat{F}_N$ is taken as an estimator of the bias of $\widehat{\theta}$ based on the distribution F. It has been proved by many authors that this approach works in many cases, that is, that it leads to consistent estimators of the properties of $\widehat{\theta}$ (e. g., Putter, 1994). The actual implementation of the bootstrap is quite simple: Drawing samples from $\widehat{F}_N$ is equivalent to drawing samples with replacement from $\{z_1, z_2, \ldots, z_N\}$.

The bootstrap is now implemented as follows: B bootstrap samples $\{z_{b1}^*, z_{b2}^*, \ldots, z_{bN}^*\}$, $b = 1, \ldots, B$, are drawn from $\widehat{F}_N$, that is, these samples are drawn with replacement from $\{z_1, z_2, \ldots, z_N\}$. From each of the B samples, the parameter $\widehat{\theta}$ is estimated, thereby obtaining B estimators θ_b^*, $b = 1, \ldots, B$. Now the expectation of θ^* (given $\widehat{F}_N$) is estimated by the mean of the estimators θ_b^*, namely, $\theta_{(.)}^* = \sum_{b=1}^B \theta_b^*/B$. The variance of θ^* (given $\widehat{F}_N$) is estimated by the variance of the estimators θ_b^*, namely, $\widehat{\text{var}}(\theta^*) = \sum_{b=1}^B (\theta_b^* - \theta_{(.)}^*)^2/B$.

The bias of $\widehat{\theta}$ is estimated by the (estimated) bias of θ^*:

$$\widehat{\text{bias}}_B = \widehat{\text{bias}}(\theta^*) = \theta_{(.)}^* - \widehat{\theta}, \tag{2.51}$$

and the bias-corrected estimator of θ is therefore

$$\begin{aligned} \widehat{\theta}_B &= \widehat{\theta} - \widehat{\text{bias}}_B \\ &= 2\widehat{\theta} - \theta_{(.)}^*. \end{aligned} \tag{2.52}$$

The variance of $\widehat{\theta}$ is simply estimated by the variance of θ_b^*:

$$\widehat{\text{var}}_B = \widehat{\text{var}}(\theta^*) = \frac{1}{B} \sum_{b=1}^{B} \left(\theta_b^* - \theta_{(.)}^*\right)^2. \tag{2.53}$$

The bootstrap as described above can also be termed the *nonparametric bootstrap*, because the distribution the bootstrap samples are drawn from is the nonparametric empirical distribution function $\widehat{F}_N$. Frequently, however, it is assumed that F is a specific distribution $F(\phi)$, only depending on a parameter (or parameter vector) ϕ, which may or may not be the same parameter as θ. Then, if ϕ is estimated by $\widehat{\phi}$, F can also be estimated by $\widetilde{F}_N = F(\widehat{\phi})$, instead of $\widehat{F}_N$. If the distributional assumption about F is correct, this *parametric* empirical distribution function will generally be a better (more efficient) estimator of F.

The *parametric bootstrap* is defined exactly analogous to the nonparametric bootstrap, except that bootstrap samples are drawn from $\widetilde{F}_N$ instead of $\widehat{F}_N$. This means that no longer samples are drawn with replacement from the original data, but from a generally more smooth distribution function. Hence, the values of the z_{bi}^* in the bootstrap sample will usually not be values also encountered in the original sample.

For example, if it is assumed that F is a normal distribution function with mean μ and variance σ^2, then bootstrap samples are drawn from a normal distribution with mean $\overline{x}$ and variance s^2, where $\overline{x}$ and s^2 are the mean and variance of the original sample.

2.10.4 Balanced bootstrap

As discussed above, the nonparametric bootstrap draws B samples of size N with replacement from the observed values $\{z_1, z_2, \ldots, z_N\}$. Taken together, these form NB drawings. Let f_i be the number of times z_i is drawn among the NB drawings. Clearly, $E(f_i) = B$. However, f_i will generally not be equal to B. This difference will lead to a nonzero estimate of the bias of an estimator, even if the estimator is unbiased. The balanced bootstrap is a resampling method that ensures that all N values z_i are drawn exactly B times in the B samples. A simple way to achieve this if memory is sufficient is to make a "supersample" consisting of B copies of the original sample, and draw B samples of size N *without* replacement from this supersample. However, more efficient algorithms to achieve the same goal are implemented in `MLA`.

In many cases, the balanced bootstrap is statistically (somewhat) more efficient than the ordinary bootstrap, especially with bias estimation. See, e.g., Davison and Hinkley (1997, section 9.2), for an extensive discussion of balanced bootstrapping.

2.10.5 Resampling regression models

Consider a simple linear regression model

$$y = \alpha + \beta x + \varepsilon,$$

where ε is a normally distributed error term with mean zero and variance σ^2. Suppose that a sample $\{(y_1, x_1), \ldots, (y_N, x_N)\}$ is available. Then parameter estimates $\widehat{\alpha}, \widehat{\beta}$, and $\widehat{\sigma}^2$ can be obtained. Now, if x is considered a random variable, nonparametric bootstrap samples can be easily obtained by resampling complete *cases*: Bootstrap samples $\{(y_1^*, x_1^*), \ldots, (y_N^*, x_N^*)\}$ consist of pairs (y_i^*, x_i^*) that are also elements of the original sample, that is, for each $i = 1, \ldots, N$, there exists a j, $1 \leq j \leq N$, such that $(y_i^*, x_i^*) = (y_j, x_j)$. Then, the parameters can be estimated from each bootstrap sample and bias-corrected estimates can be obtained, as well as an estimate of the covariance matrix of the estimator, using the formulas from section 2.10.3.

The implementation of the parametric bootstrap depends on whether a specific distribution of x is assumed. If x is regarded as a random variable with an *unspecified* distribution, the parametric bootstrap should start with drawing *nonparametric* bootstrap samples of x. If, on the other hand, a *specific* distribution of x is assumed, for example, a normal distribution with mean μ and variance σ_x^2, then the parametric bootstrap starts with drawing *parametric* bootstrap samples of x, for example, samples from a normal distribution with mean $\overline{x}$ and variance s_x^2, which are the estimates of μ and σ_x^2 from the original sample.

Given a bootstrap sample $\{x_1^*, \ldots, x_N^*\}$ of x, the parametric bootstrap draws a sample $\{\varepsilon_1^*, \ldots, \varepsilon_N^*\}$ of ε from a normal distribution with mean zero and variance $\widehat{\sigma}^2$, where $\widehat{\sigma}^2$ is the estimate of σ^2 from the original sample. Then, a bootstrap sample $\{y_1^*, \ldots, y_N^*\}$ of y is computed from the following equation:

$$y_i^* = \widehat{\alpha} + \widehat{\beta} x_i^* + \varepsilon_i^*, \tag{2.54}$$

where $\widehat{\alpha}$ and $\widehat{\beta}$ are the estimates of α and β from the original sample.

The situation is different if x is regarded as a fixed (design) variable, chosen by the experimentor. This happens, for example, if x is the dose of some drug administered to rats by the experimentor. Then each bootstrap sample should have exactly the same x values, that is, $x_i^* = x_i$ for each i in each bootstrap sample. The *parametric* bootstrap is in this case simply obtained by (2.54), with $x_i^* = x_i$. The *nonparametric* bootstrap is in this case, however, completely different from the nonparametric bootstrap with random x. In this case, first, the errors are estimated from the original sample by

$$\widehat{\varepsilon}_i = y_i - \widehat{\alpha} - \widehat{\beta}x_i. \tag{2.55}$$

Then, bootstrap samples $\{\varepsilon_1^*, \dots, \varepsilon_N^*\}$ are drawn from $\{\widehat{\varepsilon}_1, \dots, \widehat{\varepsilon}_N\}$, and bootstrap samples of y are obtained analogously to (2.54):

$$y_i^* = \widehat{\alpha} + \widehat{\beta}x_i + \varepsilon_i^*. \tag{2.56}$$

Then, bootstrap estimates of the parameters and bootstrap estimates of the covariance matrix of the parameters are obtained in the usual way (e. g., Efron, 1982, pp. 35–36).

The *jackknife* can also be implemented straightforwardly in regression models: One complete case is removed from the sample for each $\widehat{\theta}_{(i)}$ for the ungrouped jackknife, or a group of complete cases is removed for each $\widehat{\theta}_{(j)}$ for the grouped jackknife. The jackknife bias-corrected estimators and the jackknife estimators of the covariance matrix of the parameters are obtained straightforwardly (e. g., Efron, 1982, pp. 18–19).

The bootstrap and jackknife methods discussed here for regression models are the standard implementations as, for example, discussed by Efron (1982). These have some drawbacks, and therefore, alternative resampling methods have been proposed that have some advantages, for example, that they are robust to heteroskedasticity. A thorough discussion can be found in Wu (1986).

2.10.6 Resampling multilevel models

Because multilevel analysis is based on regression analysis, resampling methods for multilevel models can be based on resampling methods for regression models. The methods of section 2.10.5 can, however, not straightforwardly be applied to multilevel models, because the usual jackknife and bootstrap theory requires that the different observations be independently distributed. This is not the case with multilevel analysis, where the observations within the same Level-2 unit are dependent.

Another difference between regression analysis and multilevel analysis is that in multilevel analysis, there can be variables measured at all levels. In the two-level case, for example, there are variables describing the Level-1 units and (possibly) variables describing the Level-2 units. This implies that resampling can be performed at two levels.

Consider two-level data. A straightforward implementation of the (ungrouped) *jackknife* would be to eliminate one observation from one Level-2 unit at the time to obtain a jackknife sample. This resampling scheme is exactly equivalent to the resampling scheme of the standard ungrouped jackknife of section 2.10.1. Another possibility is to implement the grouped jackknife. With the grouped jackknife, it is most logical to use the Level-2 units as groups. The Level-2 units may have different sizes, and therefore, the grouped jackknife with unequal group sizes should be used.

The *parametric bootstrap* can be easily implemented in multilevel analysis. If the X_j and W_j variables are considered fixed in (2.1) and (2.2), bootstrap samples $\{y_{b1}^*, \dots, y_{bJ}^*\}$ can be obtained in the following way. First, for each $j = 1, \dots, J$, draw a bootstrap Level-2 error vector u_j^* from a normal distribution with mean zero and covariance matrix $\widehat{\Theta}$. Then, draw a bootstrap Level-1 error vector ε_j^* from a normal distribution with mean zero and covariance matrix $\widehat{\sigma}_\varepsilon^2 I_{N_j}$. Finally, the bootstrap sample of y is obtained from

$$\beta_j^* = W_j\widehat{\gamma} + u_j^* \tag{2.57}$$

and

$$y_j^* = X_j\beta_j^* + \varepsilon_j^*. \tag{2.58}$$

Then, bias-corrected bootstrap estimators and bootstrap estimators of the covariance matrix of the parameters are obtained in the usual way. This is the parametric bootstrap that is implemented in MLA. It is also possible to derive a parametric bootstrap estimator in case the X and W variables are considered random. This is analogous to (2.54), but it is not implemented in MLA.

For the *nonparametric bootstrap*, several situations can be studied. If the X and W variables can be considered *fixed*, then, analogously to regression analysis, the *errors* have to be estimated. As explained in section 2.7, the shrunken residuals (2.32) and (2.34) can be used as estimators of the Level-2 and Level-1 errors, respectively. A drawback of these errors may be that their variances are less than the variances in the population. When, however, sample sizes at both levels increase, this difference diminishes. But, alternatively, the *raw residuals* (2.30)–(2.31) can be used instead of the shrunken residuals.

Unlike in regression analysis, the estimated residuals in multilevel analysis do not necessarily have a zero mean. Therefore, the means are subtracted first. Otherwise, the possibly nonzero mean of the errors would necessarily lead to biased estimators of the constant. Once (centered) estimates $\{\widehat{u}_j\}$, $j = 1, \ldots, J$, and $\{\widehat{\varepsilon}_{ij}\}$, $j = 1, \ldots, J$, $i = 1, \ldots, N_j$, of the errors are obtained, nonparametric bootstrap samples $\{u_j^*\}$, $j = 1, \ldots, J$, and $\{\varepsilon_{ij}^*\}$, $j = 1, \ldots, J$, $i = 1, \ldots, N_j$ are drawn, and nonparametric bootstrap samples of y are obtained from (2.57) and (2.58). Then, estimators can be obtained in the usual way, and bootstrap bias-corrected estimators and standard errors can be obtained straightforwardly. This bootstrap procedure of resampling from estimated errors is called the *error bootstrap*.

If the X and W variables are considered *random*, *nonparametric* bootstrap samples can be drawn by resampling complete *cases*. This is, however, somewhat more complicated than in regression analysis, because the hierarchical structure of the data should be respected. The bootstrap samples can be drawn in the following way. First, a sample of size J is drawn with replacement from the *Level-2 units*. This gives a sample j_k^*, $k = 1, \ldots, J$ of Level-2 unit numbers and accompanying Level-2 variables $W_{j_k^*}$. Then for each k, a nonparametric bootstrap sample of complete cases from the (original) unit $j = j_k^*$ is drawn, giving $\{(y_{ik}^*, X_{ik}^*)$, $k = 1, \ldots, J, i = 1, \ldots, N_{j_k^*}\}$. This is called the *cases bootstrap* for both levels.

It is also possible to draw bootstrap samples from the Level-2 units only, keeping all the y's, X's, and W's fixed once a Level-2 unit is drawn. This is useful when the data within the unit can not be considered a simple random sample, for example, with repeated measures data or families. Then, a complete Level-2 unit is (temporarily) regarded as a single observation and bootstrap samples are drawn from these observations. With repeated measures, this implies that for each subject that is drawn in the bootstrap sample, the data for all the timepoints are exactly the same as in the original sample. For a family, this means that the complete family is kept together, and that, once the family is drawn in the bootstrap sample, mother, father, and children are all part of the bootstrap sample, and, for example, the mother can not be drawn twice within the same Level-2 unit.

On the other hand, it is also possible to keep the Level-2 units fixed, and draw bootstrap samples only from the Level-1 units within each Level-2 unit. This can be useful when the Level-2 units can not be considered a simple random sample, for example, when several (prespecified) countries are compared and people within each country are drawn randomly. Then, in the bootstrap samples, all countries are present once, just as in the original sample. Bootstrap samples are drawn from complete cases within each country.

Once bootstrap samples are drawn, bootstrap bias-corrected estimators and bootstrap standard errors can be obtained straightforwardly.

2.10.7 Bootstrap confidence intervals

Up till now, we have used the bootstrap only for bias correction and computation of standard errors. However, an important and nontrivial application of the bootstrap is the computation of confidence intervals. We will now discuss a number of different types of bootstrap confidence intervals for a typical parameter θ with true value θ_0. We will only discuss two-sided intervals, one-sided intervals are defined analogously. The intended nominal coverage of the confidence interval will be denoted by $1-\alpha$, so that the probability that the interval contains the true parameter value should be approximately $1-\alpha$.

Notation Before we introduce the different bootstrap confidence intervals, we will introduce some useful notation. Let $\Phi(z)$ be the standard normal distribution function. Then z_α is the α-th quantile of the standard normal distribution, $z_\alpha = \Phi^{-1}(\alpha)$. Let the distribution function of the estimator $\widehat{\theta}$ be $H(\theta)$, that is, $H(\theta) = \Pr(\widehat{\theta} \le \theta)$. A consistent estimator of this distribution function is obtained from the B bootstrap replications θ_b^*, $b = 1, \ldots, B$, of $\widehat{\theta}$:

$$\widehat{H}(\theta) = \frac{\#\{b \,:\, \theta_b^* \le \theta\}}{B}\,. \tag{2.59}$$

Note that $\widehat{H}$ is invariant under monotonic transformation, in the sense that if $g(\theta)$ is a monotonically increasing function of θ, then the estimate of its distribution function is

$$\tilde{H}(g(\theta)) = \frac{\#\{b \,:\, g(\theta_b^*) \le g(\theta)\}}{B} = \widehat{H}(\theta)\,.$$

This property has been used in the derivations of some of the confidence intervals described below.

Bootstrap normal confidence interval If the assumptions of the model, including the normality assumptions, hold, then the estimators are asymptotically normally distributed with a certain covariance matrix, derived from the likelihood function. Hence, for our typical parameter θ, we have

$$\sqrt{N}(\widehat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \psi), \tag{2.60}$$

say. The distribution of $\widehat{\theta} - \theta_0$ can be approximated by the normal distribution with mean zero and variance $\widehat{\psi}/N$, where $\widehat{\psi}$ is a consistent estimator of ψ derived from the likelihood function. The usual confidence intervals for θ_0 are therefore

$$\left[\widehat{\theta} + z_{\frac{1}{2}\alpha}\widehat{\text{se}}_N(\widehat{\theta});\ \widehat{\theta} + z_{1-\frac{1}{2}\alpha}\widehat{\text{se}}_N(\widehat{\theta})\right], \tag{2.61}$$

where $\widehat{\text{se}}_N(\widehat{\theta}) = \sqrt{\widehat{\psi}/N}$ is the estimator of the asymptotic standard deviation of $\widehat{\theta}$. Under mild regularity conditions, the estimators are asymptotically normally distributed, even if the random terms in the model are not. In that case, $\widehat{\text{se}}_N$ may not be a consistent estimator of the standard deviation of the estimators of the variance components, although it is still consistent for the fixed parameters. This suggests replacing $\widehat{\text{se}}_N$ in (2.61) by a bootstrap estimator. This gives the *bootstrap normal* confidence interval

$$\left[\widehat{\theta} + z_{\frac{1}{2}\alpha}\widehat{\text{se}}_B(\widehat{\theta});\ \widehat{\theta} + z_{1-\frac{1}{2}\alpha}\widehat{\text{se}}_B(\widehat{\theta})\right], \tag{2.62}$$

in which $\widehat{\text{se}}_B$ is the bootstrap estimator of the standard deviation of $\widehat{\theta}$. Alternatively, one might use

$$\left[\widehat{\theta}_B + z_{\frac{1}{2}\alpha}\widehat{\text{se}}_B(\widehat{\theta});\ \widehat{\theta}_B + z_{1-\frac{1}{2}\alpha}\widehat{\text{se}}_B(\widehat{\theta})\right], \tag{2.63}$$

where $\widehat{\theta}_B$ is the bootstrap bias-corrected estimator of θ.

The bootstrap normal confidence interval relaxes the assumption of normality of the data, but still heavily relies on the asymptotic normality of the estimators. In finite samples, however, the estimators may not be approximately normally distributed (Busing, 1993).

Hall's percentile interval Hall's percentile interval (Hall, 1992, p. 12) takes the bootstrap normal interval (2.62) as its starting point. That interval is based on the idea that

$$\Pr\left(\widehat{\theta} + z_{\frac{1}{2}\alpha}\widehat{\text{se}}_B(\widehat{\theta}) \leq \theta_0 \leq \widehat{\theta} + z_{1-\frac{1}{2}\alpha}\widehat{\text{se}}_B(\widehat{\theta})\right) \longrightarrow 1-\alpha, \tag{2.64}$$

because $\widehat{\theta}$ is asymptotically normally distributed and $\widehat{\text{se}}_B(\widehat{\theta})$ is a consistent estimator of its standard deviation. In finite samples, however, the distribution of $\widehat{\theta}$ may not be approximately normal (Busing, 1993). Therefore, instead of using quantiles of the normal distribution, using bootstrap quantiles may give more accurate results.

To derive the necessary bootstrap quantiles, let us rewrite (2.64) into the following form:

$$\Pr\left(z_{\frac{1}{2}\alpha}\widehat{\text{se}}_B(\widehat{\theta}) \leq \theta_0 - \widehat{\theta} \leq z_{1-\frac{1}{2}\alpha}\widehat{\text{se}}_B(\widehat{\theta})\right) \longrightarrow 1-\alpha.$$

The estimated quantiles $q_{\frac{1}{2}\alpha} = z_{\frac{1}{2}\alpha}\widehat{\text{se}}_B(\widehat{\theta})$ and $q_{1-\frac{1}{2}\alpha} = z_{1-\frac{1}{2}\alpha}\widehat{\text{se}}_B(\widehat{\theta})$ of the normal distribution have to be replaced by quantiles of the distribution of $\theta_0 - \widehat{\theta}$. These are estimated by quantiles $\widehat{q}_{\frac{1}{2}\alpha}$ and $\widehat{q}_{1-\frac{1}{2}\alpha}$ of the bootstrap distribution of $\widehat{\theta} - \theta^*$. From the definition $\Pr(\widehat{\theta} - \theta^* \leq \widehat{q}_{\frac{1}{2}\alpha}) = \frac{1}{2}\alpha$, it follows that $\widehat{q}_{\frac{1}{2}\alpha} = \widehat{\theta} - \widehat{H}^{-1}(1-\frac{1}{2}\alpha)$ and the confidence interval for θ_0 becomes $[\widehat{\theta} + \widehat{q}_{\frac{1}{2}\alpha}; \widehat{\theta} + \widehat{q}_{1-\frac{1}{2}\alpha}]$, which reduces to

$$\left[2\widehat{\theta} - \widehat{H}^{-1}(1-\tfrac{1}{2}\alpha);\ 2\widehat{\theta} - \widehat{H}^{-1}(\tfrac{1}{2}\alpha)\right]. \tag{2.65}$$

Note that the upper quantile of $\widehat{H}$ ends up (in reverse) in the lower confidence point and vice versa. This tends to give a small bias and skewness correction.

This percentile interval has not been implemented in `MLA`, because the percentile-*t* interval is generally considered to be an improvement of it.

Percentile-*t* The percentile-*t* (also called bootstrap-*t*) is a combination of the ideas of the bootstrap normal and Hall's percentile intervals. It is derived by rewriting (2.64) into the following form:

$$\Pr\left(z_{\frac{1}{2}\alpha} \leq \frac{\theta_0 - \widehat{\theta}}{\widehat{\text{se}}_B(\widehat{\theta})} \leq z_{1-\frac{1}{2}\alpha}\right) \longrightarrow 1-\alpha. \tag{2.66}$$

The quantiles of the normal distribution are now replaced by quantiles of the distribution of $\widehat{t} = (\theta_0 - \widehat{\theta})/\widehat{\text{se}}_B(\widehat{\theta})$. These are estimated by quantiles of the bootstrap distribution of $t^* = (\widehat{\theta} - \theta^*)/\text{se}^*_B(\theta^*)$. Let $\widehat{G}(t)$ be the bootstrap-estimated distribution function of this quantity, i.e.,

$$\widehat{G}(t) = \frac{\#\left\{b\ :\ \dfrac{\widehat{\theta} - \theta^*_b}{\text{se}^*_{B,b}(\theta^*)} \leq t\right\}}{B},$$

and let $\widehat{t}_{\frac{1}{2}\alpha}$ and $\widehat{t}_{1-\frac{1}{2}\alpha}$ be the $\frac{1}{2}\alpha$-th and $(1-\frac{1}{2}\alpha)$-th quantiles of $\widehat{G}$, respectively, that is, $\widehat{t}_{\frac{1}{2}\alpha} = \widehat{G}^{-1}(\frac{1}{2}\alpha)$ and $\widehat{t}_{1-\frac{1}{2}\alpha} = \widehat{G}^{-1}(1-\frac{1}{2}\alpha)$. The percentile-$t$ interval is obtained by replacing $z_{\frac{1}{2}\alpha}$ by $\widehat{t}_{\frac{1}{2}\alpha}$ and $z_{1-\frac{1}{2}\alpha}$ by $\widehat{t}_{1-\frac{1}{2}\alpha}$ in (2.62) and is thus

$$\left[\widehat{\theta} + \widehat{t}_{\frac{1}{2}\alpha}\widehat{\text{se}}_B(\widehat{\theta});\ \ \widehat{\theta} + \widehat{t}_{1-\frac{1}{2}\alpha}\widehat{\text{se}}_B(\widehat{\theta})\right]. \tag{2.67}$$

This confidence interval requires an estimate $\text{se}^*_{B,b}(\theta^*)$ of the standard deviation of θ^* for each bootstrap resample b. This is usually obtained by performing a small bootstrap within each bootstrap resample. Thus, for example, $B = 1000$ bootstrap samples are drawn with replacement from the original sample and within each sample $b = 1, \ldots, B$, $B_2 = 25$ samples are drawn with replacement from the bootstrap sample. From the B_2 samples, $\text{se}^*_{B,b}(\theta^*)$ is obtained. This means that $B \cdot B_2$ bootstrap samples have to be drawn and $B \cdot B_2$ times the estimator of θ has to be computed. In the example, this amounts to $1000 \cdot 25 = 25\,000$ bootstrap samples and 25 000 times computing the estimator.

The percentile-t interval tends to perform better than the bootstrap normal and Hall's percentile interval, because it uses the nonnormality of the distribution of the estimator (as opposed to the former) and $\widehat{t}$ is more nearly *pivotal* than $\theta_0 - \widehat{\theta}$ in a number of important cases, which means that its distribution depends less on the parameters that are being estimated. The quantity $\widehat{t}$ is not always nearly pivotal, however, and in those cases in which it is not, the percentile-t confidence interval performs less well. A complicated extension that aims at transforming the parameter to a near-pivotal quantity is the *variance stabilized percentile-t* interval, see, e.g., Efron and Tibshirani (1993, section 12.6).

Efron's percentile interval The idea behind this interval is quite different from the ideas behind the bootstrap normal interval and its extensions. It was stated above that $\widehat{H}(\theta)$ is a consistent estimator of the distribution function of $\widehat{\theta}$. Therefore, an asymptotic $1-\alpha$ confidence interval can be obtained by taking the relevant quantiles from $\widehat{H}$, which leads to the interval

$$\left[\widehat{H}^{-1}(\tfrac{1}{2}\alpha);\ \ \widehat{H}^{-1}(1-\tfrac{1}{2}\alpha)\right]. \tag{2.68}$$

Efron's percentile interval does not rely on the asymptotic normality of $\widehat{\theta}$. Its coverage performance in finite samples is, however, frequently not very well, because the end points of the interval tend to be a little biased. Note the difference with Hall's percentile interval. Here, percentiles of the distribution of $\widehat{\theta}$ are approximated by percentiles of the distribution of θ^*, whereas in Hall's percentile interval, percentiles of the distribution of $\theta_0 - \widehat{\theta}$ are approximated by percentiles of the distribution of $\widehat{\theta} - \theta^*$.

Efron's percentile interval is simply called "percentile" in the input and output of `MLA`.

Bias-corrected (BC) and bias-corrected and accelerated (BC$_a$) percentile intervals The BC and BC$_a$ intervals have been introduced to correct for some bias in the endpoints of Efron's percentile interval (2.68). Assume that there exists a monotonically increasing function $g(\theta)$ such that

$$\frac{g(\widehat{\theta}) - g(\theta_0)}{1 + ag(\theta_0)} \sim \mathcal{N}(-z_0, 1). \tag{2.69}$$

The constant z_0 allows for some bias in the estimator $g(\widehat{\theta})$ of $g(\theta_0)$ and the constant a, called the *acceleration constant*, expresses the speed at which the standard deviation of the estimator increases with the parameter being estimated. In typical estimation problems, $a = O(N^{-1/2})$ and $z_0 = O(N^{-1/2})$.

From the likelihood based on (2.69), it can now be derived that the exact confidence interval for θ_0 is equal, up to order $O(N^{-1})$, to the BC$_a$ interval given by

$$\left[\widehat{H}^{-1}\left(\Phi(z[\tfrac{1}{2}\alpha])\right); \widehat{H}^{-1}\left(\Phi(z[1-\tfrac{1}{2}\alpha])\right)\right], \tag{2.70}$$

where

$$z[\tfrac{1}{2}\alpha] = z_0 + \frac{z_0 + z_{\frac{1}{2}\alpha}}{1 - a(z_0 + z_{\frac{1}{2}\alpha})}$$

and $z[1-\frac{1}{2}\alpha]$ similarly defined. Note that this interval does not depend on the specific transformation $g(\cdot)$, which follows from the invariance property of $\widehat{H}$ discussed earlier. In practice, the constants z_0 and a have to be estimated, but this does not alter the results up to order $O(N^{-1})$. Moreover, even if (2.69) does not hold, the BC$_a$ endpoints are correct up to order $O(N^{-1})$, whereas in many cases the endpoints of the intervals discussed previously are only correct up to order $O(N^{-1/2})$.

A simple consistent estimator of z_0 is $\widehat{z}_0 = \Phi^{-1}\left[\widehat{H}(\widehat{\theta})\right]$. The estimation of a is the most important problem with the BC$_a$ method. If it is assumed that $a = 0$, we obtain the BC interval, which is discussed, e.g., in Efron (1982). Usually, however, the BC interval is only correct up to order $O(N^{-1/2})$ and is therefore typically worse than the BC$_a$ interval.

Efron (1987) provided several formulas for a. In a one-parameter parametric model where $\widehat{\theta}$ is the ML estimator, a good approximation for a is $a \approx \frac{1}{6}\text{SKEW}_{\theta=\widehat{\theta}}(\dot{l}_\theta)$, where SKEW denotes the skewness of a random variable and $\dot{l}_\theta$ is the score function (derivative of the loglikelihood with respect to θ).

When more parameters are to be estimated, which is the case in multilevel analysis, these results are no longer valid. Efron (1987) gave a formula for a based on reducing the multiparameter problem to a one-parameter problem defined by the so-called *least favorable direction*. This formula may be used for the parametric bootstrap in multilevel analysis, although the expressions will be quite complicated. This was done for a simple two-level variance components model by LeBlond (2005, sec. 7.3.4), but expressions for general implementation in MLA are not yet available.

For the nonparametric bootstrap, the standard formula for a is based on the empirical influence function of $\widehat{\theta}$. This is, however, not well defined for multilevel data, so that this formula cannot be used. Tu and Zhang (1992) proposed to estimate a by the jackknife according to the formula

$$\widehat{a}_J = \frac{(N-1)^3}{6N^3(\widehat{\sigma}^2_{J(1)})^{3/2}} \sum_{i=1}^{N} \left(\widehat{\theta}_{(i)} - \bar{\theta}_{(1)}\right)^3. \tag{2.71}$$

For multilevel data, we would have to replace this jackknife formula with a grouped jackknife method for unequal group sizes (see section 2.10.1). It is, however, doubtful whether the jackknife for multilevel models will give a reasonable estimate of a third-order moment. A bootstrap analog of (2.71) would be

$$\widehat{a}_B = \frac{1}{6}\frac{\dfrac{1}{B}\displaystyle\sum_{b=1}^{B}\left(\theta^*_b - \theta^*_{(.)}\right)^3}{\left[\dfrac{1}{B}\displaystyle\sum_{b=1}^{B}\left(\theta^*_b - \theta^*_{(.)}\right)^2\right]^{3/2}}.$$

It is still an open question whether this gives reasonable results.

The BC interval is available in `MLA`, but due to the problems with estimating the acceleration constant, the BC$_a$ is not yet available.

2.11 Effective sample size and intra-class correlation

The intra-class dependency in multilevel data generally affects the precision of the estimators of the (fixed) parameters. This means that, in comparison with an independent sample with the same variance of the (total) error term, a larger or smaller sample is needed to achieve the same precision. Equivalently, the precision obtained from a sample containing N Level-1 units with multilevel data is equal to the precision from a sample of, say, N' observations with independent data. We can then say that the *effective sample size* of the multilevel dataset is only N', whereas actual sample size is N. This terminology is due to Kish (1965, p. 259). Here, we will analyze this in deeper detail.

First, assume that the data are drawn from a two-level random effects ANOVA model without explanatory variables,

$$y_{ij} = \gamma + u_j + \varepsilon_{ij},$$

where γ is the grand mean, which is equal to the intercept in this case, and all other notation is obvious. Let Θ be the variance of the Level-2 random effect and σ_ε^2 be the variance of the Level-1 error term. Then $\sigma_y^2 = \Theta + \sigma_\varepsilon^2$ is the variance of y_{ij} and $\rho = \Theta/(\sigma^2 + \Theta)$ is the intra-class correlation. If the sample sizes in all Level-2 units are equal, i.e., $N_j = N/J = N^*$, say, for all j, then the variance of the sample mean $\bar{y}$ of y is

$$E\left[(\bar{y} - \gamma)^2\right] = \frac{1}{N}(\sigma_\varepsilon^2 + N^*\Theta) = \frac{\sigma_y^2[1 + (N^* - 1)\rho]}{N}.$$

On the other hand, the variance of the sample mean of an independently distributed sample with mean γ and variance σ_y^2 is σ_y^2/N. Consequently, the variance of $\bar{y}$ based on a multilevel sample of size N is equal to the variance of $\bar{y}$ based on an independent sample of size $N' = N/[1 + (N^* - 1)\rho]$ and this expression for N' is the effective sample size. Hox (2002, p. 5) also correctly gives this formula for the effective sample size, but incorrectly states that the standard error is multiplied by the factor $1 + (N^* - 1)\rho$, whereas our analysis shows that it is the variance that is multiplied by this factor. Because $\rho > 0$ and $N^* > 1$, the denominator is larger than 1 and thus the effective sample size is smaller than N.

If not all sample sizes in the Level-2 units are equal, then the variance of the sample mean $\bar{y}$ of y is

$$\frac{\sigma_y^2}{N}\left[1 + \left(\sum_{j=1}^{J}\frac{N_j^2}{N} - 1\right)\rho\right],$$

so if γ is estimated by the sample mean, then the effective sample size is

$$\frac{N}{\left[1 + \left(\sum_{j=1}^{J}\frac{N_j^2}{N} - 1\right)\rho\right]}.$$

However, unlike the case of equal group sizes, the unweighted sample mean is not the most efficient estimator anymore. If Θ and σ_ε^2 are known, the most efficient estimator is the GLS estimator $\widehat{\gamma} = (X'V^{-1}X)^{-1}X'V^{-1}y$, where y is the column vector consisting of all y_{ij}'s, X is the matrix of explanatory variables corresponding to the fixed coefficients, which in this case is a column vector consisting only of ones, and V is a diagonal block matrix consisting of blocks $V_j = Z_j \Theta Z_j' + \sigma_\varepsilon^2 I_{N_j}$, where Z_j is a matrix of explanatory variables corresponding to the random coefficients, which in this case is also a column vector (but of length N_j instead of N). The variance of $\widehat{\gamma}$ is

$$(X'V^{-1}X)^{-1} = \frac{\sigma_y^2}{\sum_{j=1}^{J} \frac{N_j}{1+(N_j-1)\rho}}$$

and thus, the effective sample size is

$$\sum_{j=1}^{J} \frac{N_j}{1+(N_j-1)\rho}.$$

Usually, of course, Θ and σ_ε^2 are not known and must be estimated. However, this does not influence the (asymptotic) variance of the GLS estimator. The `FIML` estimator and two-step ML estimator based on the `REML` estimator are of this form. In any case, the effective sample size is smaller than the actual sample size.

Now, consider the following simple random effects ANCOVA model:

$$y_{ij} = \gamma_1 + \gamma_2 x_{ij} + u_j + \varepsilon_{ij},$$

which is the random effects ANOVA discussed above, extended with the covariate x_{ij}, which has a fixed coefficient. The matrix X is now $N \times 2$, with the first column consisting of ones and the second consisting of the observations x_{ij}. The covariance matrix V is unaltered, and the (asymptotic) covariance matrix of the GLS (`FIML`, two-step ML) estimators of the fixed coefficients $\gamma = (\gamma_1, \gamma_2)'$ is still given by the expression $(X'V^{-1}X)^{-1}$. Furthermore, assume that all within-group sample sizes are equal ($N_j = N^* = N/J$ for all j) and even, and $x_{ij} = 1$ if i is odd and $x_{ij} = -1$ if i is even. Then

$$(X'V^{-1}X)^{-1} = \frac{\sigma_r^2}{N} \begin{pmatrix} 1+(N^*-1)\rho & 0 \\ 0 & 1-\rho \end{pmatrix},$$

where $\sigma_r^2 = \Theta + \sigma_\varepsilon^2$ is the variance of the random part. With independent data with residual variance σ_r^2, the variance of the optimal estimator of γ, which is the `OLS` estimator, is $\sigma_r^2 I_2/N$. Consequently, the effective sample size for γ_1 is $N/[1+(N^*-1)\rho]$ as before, but the effective sample size for γ_2 is $N/(1-\rho)$, which is larger than N.

Explanatory variables with random coefficients complicate the situation even further. As shown in (2.4), such models can be written as mixed linear models. For a typical observation, this is written as

$$y_{ij} = x_{ij}'\gamma + z_{ij}'u_j + \varepsilon_{ij},$$

where x_{ij} and z_{ij} are vectors of explanatory variables (typically including the constant), γ is a vector of fixed coefficients, u_j is a vector of random coefficients with mean zero, and ε_{ij} is a random residual. Let, as usual, Θ be the covariance matrix of u_j, and let $r_{ij} = z_{ij}'u_j + \varepsilon_{ij}$ be the total residual. Then the variance of r_{ij} is $z_{ij}'\Theta z_{ij} + \sigma_\varepsilon^2$, which generally depends on z_{ij}. Analogously, the covariance of the residuals of observations ij and kj is $z_{ij}'\Theta z_{kj}$ and the intra-class correlation of these two observations is

$$\rho_{ij,kj} = \frac{z_{ij}'\Theta z_{kj}}{(z_{ij}'\Theta z_{ij} + \sigma_\varepsilon^2)^{1/2}(z_{kj}'\Theta z_{kj} + \sigma_\varepsilon^2)^{1/2}}.$$

Thus, this depends on z_{ij} and z_{kj}. For this reason, it may be called the *conditional intra-class correlation*, given the values z_{ij} and z_{kj}. In MLA, the conditional intra-class correlation is computed for the (hypothetical) case where all elements of z_{ij} and z_{kj} are zero, except for the constant. Then the conditional intra-class correlation has the familiar form

$$\rho_0 = \frac{\Theta_{11}}{\Theta_{11} + \sigma_\varepsilon^2},$$

where Θ_{11} is the (diagonal) element of Θ corresponding to the constant. Whether ρ_0 is an interesting value depends on the average and variability of z. It is most relevant if z is centered around the grand mean (except for the constant, of course) and it does not vary much. Goldstein, Browne, and Rasbash (2002, p. 225) study $\rho_{ij,kj}$ for the case $z_{ij} = z_{kj}$ and call it *variance partition coefficient* (VPC). They show that it may vary strongly with z_{ij}, so a single value like ρ_0 may not be very relevant.

The intra-class correlation depends on the values of the explanatory variables and so does the variance of the (total) residual. We cannot expect to find a simple expression for the effective sample size involving these characteristics if these do not themselves have a simple expression that is the same for all observations. However, we can compute a kind of effective sample size in the following way. Let ψ_{1k} be the (asymptotic) variance of the (most efficient) estimator of a fixed parameter γ_k if the residuals have the same variances as in the multilevel model, but are all independent, and let ψ_{2k} be the (asymptotic) variance of the (most efficient) estimator of γ_k with the intra-class dependency present. With these definitions, the effective sample sizes in the cases studied above can all be written as

$$\mathrm{ESS}_k = N\frac{\psi_{2k}}{\psi_{1k}}. \tag{2.72}$$

As stated above and proved in section A.4, ψ_{2k} is the k-th diagonal element of the matrix $(X'V^{-1}X)^{-1}$. Similarly, ψ_{1k} is given by the same formula, but with V replaced by the matrix Ω that has the same diagonal elements as V, but has all its off-diagonal elements equal to zero. With these values of ψ_{1k} and ψ_{2k}, we define (2.72) as the effective sample size for the k-th fixed parameter.

Note that we have only considered effective sample sizes for the fixed parameters. If one is interested in the random parameters, it does not make much sense to make comparisons with an independent sample, because the random parameters are mainly substantively interesting for modeling the intra-class dependencies. Therefore, effective sample sizes for these are not particularly interesting. However, most random parameters can be estimated consistently from an independent sample, because they induce a certain parametric form of heteroskedasticity. Only the Level-1 variance σ_ε^2 and the Level-2 variance Θ_{11} corresponding with the constant cannot be disentangled. Only their sum can be estimated consistently from an independent sample. So for all other random parameters, an effective sample size could be computed, but this is not done in MLA because it is not very interesting, because it is considerably more complicated, and because it relies heavily on strong assumptions like normality of the residuals and homoskedasticity at Level-1.

2.12 Missing data

Missing data are a frequently occurring phenomenon. For instance, in repeated measures designs, the points in time at which the different subjects are measured may not be the same, or the number of points in time the subjects are measured may differ. This situation leads to missing time-points, that is, all time-specific variables of a subject are missing at some point in time. However, the time-invariant variables (such as sex) are, of course, known. This situation is easily handled by a multilevel model, in which the subjects are the Level-2 units, and the time-points are the Level-1 units. As was discussed for a usual multilevel model, the number of Level-1 units may be different for different Level-2 units, and so the missing timepoints give no problems. An example of repeated measures is given in chapter 4.

If, however, in a multilevel model, be it an application in repeated measures or not, for some Level-1 unit, some Level-1 variables are measured, but others are not (or for some Level-2 unit, some Level-2 variables are measured, but others are not), there are missing values that can not be handled by the standard model. If only output variables are missing, the EM algorithm provides a standard way of dealing with the missing values in a satisfactory way. If, however, some exogenous (X and/or Z) variables are missing, the EM algorithm can not be used straightforwardly, because it requires that the joint distribution of the exogenous and the endogenous (output) variables is known. Standard multilevel modeling only assumes that the conditional distribution of the output variables given the exogenous variables is known. This poses severe complications.

If the amount of data that is missing is relatively small, standard ad-hoc solutions to the missing-data problem can be used, such as *listwise deletion* (deletion of cases with one or more missing values), *pairwise deletion* (computation of “sufficient” statistics, such as covariances, on the basis of all available information for the variables in question), *mean substitution* (substitution of the mean of the observed values of a variable for a missing value on that variable), or other substitution methods. All these methods have their advantages and drawbacks and none is fully satisfactory, especially when the number of missing values is large.

In the current version of MLA, no specific means of missing-data handling are implemented. Listwise deletion and several forms of substitution can be done by the user before the data set is processed by MLA. Pairwise deletion can not be done, because the program requires raw data. In principle, pairwise deletion could be done within the program, but this is not implemented (yet).

Chapter 3

Commands

MLA runs as a stand-alone batch program. It uses an input file and an output file as parameters. This means that the program can be started by the command

```
MLA  [-hHv]  input-file-name  output-file-name
```

where *input-file-name* is the name of the file that contains the input and *output-file-name* is the name of the file in which the output of the program will be saved. Both files are simple text files (ascii). The output file will be explained in the next chapter. The input file will be considered here. The options are help (-h), extended help (-H) and verbosity (-v), respectively. The latter shows some information on the terminal about what the program is doing.

The input file consists of statements, which are case insensitive. Every statement begins with a slash and a keyword (e. g., /TITLE). Every keyword may be abbreviated, but it must be at least of length three to be recognized (e. g., /TIT). Other text following the keyword and/or leading spaces will be ignored. The rest of the statements must follow on lines below the keyword and should precede the next statement. These lines are called substatements and may also consist of one or more keywords (e. g., file). The last statement to be read is the /END statement. All other statements, and corresponding substatements, may appear in any order (but before the /END statement if they are to be executed). A substatement may continue on the next line. In this case the first line must be ended with two backslashes (\\). Finally, comments, preceded by a percent sign (%), may appear throughout the input file. All text on a line, after and including the percent sign, will serve as comment and is ignored as program input.

In the following, all statements and substatements implemented are discussed and illustrated with small examples. In Chapter 4, where we focus on the program output, complete examples are provided.

3.1 /TITLE (optional)

Following the keyword /TITLE, the first non-blank line contains the title for the analysis. Although the statement is optional, it is highly recommended. Moments after the analysis all may seem clear, but after a few months you may have no idea what you have done. The title may be your only clue. You may also enrich your input file with comments. In contrast to comments, the title is repeated on top of every part of the output.

Example:

```
/TITLE
  MLA example 1: analysis of variance
```

3.2 /DATA (required)

The /DATA statement contains information about the data file. This statement has seven substatements, three of which are required. The file substatement gives the name of the data file, variables the number of variables in the data file, id1 the (optional) variable number of the Level-1 identifier variable, and id2 the variable number of the Level-2 identifier variable. The missing substatement specifies which value of a variable indicates a missing value, and centering and level-2 centering are options for centering the data before further analysis.

Example:

```
/DATA
  file      = sesame.dat  % data set from Glasnapp and Poggio (1985)
  variables = 3           % total of three variables
  id2       = 1           % Level-2 identification given by first variable
  missing   = v3(999)     % v3=999 means missing
```

3.2.1 file (required)

This substatement indicates the name of the data file. The name is given after the equals sign and must satisfy the usual DOS conventions on filenames. If the file is in the current directory the complete pathname is not necessary. The file itself is a free-field formatted numbers-only ascii file. This means that values of variables must be separated by at least one blank. A case may consist of more than one line. Cases must be sorted by the Level-2 identifier variable (see below).

3.2.2 variables (required)

The variables substatement specifies the number of variables in the data file. Because the data file is a free-field formatted file and one case may consist of more than one line, this is necessary information for the program to determine when to start a new case.

3.2.3 id1 (optional)

With this substatement, a case number variable can be given. Level-1 units are interchangeable within a Level-2 unit. Therefore, a Level-1 identifier variable is not necessary. However, it can be useful in those situations where the output gives specific information about cases at the first level. The variable number has to follow the keyword id2 and it must indicate the position of the identifier variable in the data file. The variable number must be at least 1 and less than or equal to the number of variables, indicated in the variables substatement. If this substatement is omitted, the order in which the Level-1 units are read from the data file is used as identification.

3.2.4 id2 (required)

One of the variables in the data file must contain a code (number) that identifies the Level-2 units. This may be a group number or, in case of repeated measurements, a subject number. The number is essential for a correct discrimination of the Level-2 units. The variable number has to follow the keyword id2 and it must indicate the position of the identifier variable in the data file. The variable number must be at least 1 and less than or equal to the number of variables, indicated in the variables substatement.

3.2.5 `missing` (optional)

For every variable, one missing value may be specified on this substatement. After the equal sign, first the variable is indicated followed by the missing value between parentheses. More variables and values are separated by commas.

3.2.6 `centering` (optional)

If this substatement is given, the grand mean of the variables is subtracted. Following the `centering` substatement, the variable numbers of the variables that will be centered are given, separated by commas. These variables will be centered (ignoring grouping) directly after reading the data, but before any analysis.

A warning must be given here. The choice whether to center, and if so, how to center, is an intricate one and may lead to unexpected results. It is much more complicated than with non-hierarchical data. This has led to extensive discussions in the multilevel modeling literature. See section 2.3 for an introduction to this topic and references to this literature.

3.2.7 `level-2 centering` (optional)

Level-2 centering means centering within contexts, i.e., subtracting the group mean. Following the `level-2 centering` substatement, the variable numbers of the variables that will be centered are given, separated by commas. These variables will be centered within groups directly after reading the data, but before any analysis. As mentioned above, the choice to center may lead to unexpected results. See section 2.3 for a discussion.

3.3 `/MODEL` (required)

The `/MODEL` statement is followed by a set of equations that specify the model that has to be estimated. Every equation must be on a single line. There is only one Level-1 equation, but there may be one or more Level-2 equations. The order in which the Level-1 and Level-2 equations appear is arbitrary. The terms used in the Level-1 equation are:

- `V`*i* = variable i, which is the i-th variable in the data file. `V`*i* may be either indicating the outcome variable or a predictor variable.

- `B`*k* = beta component k, i.e., β_{jk}, the k-th element of a typical β_j, cf. (2.1). At Level-1 these are the random regression coefficients, which are the outcome variables at Level-2, cf. (2.2).

- `E` = the Level-1 random term (ε). This term is considered to be a residual or error term. The variance of this term has to be estimated from the data.

The Level-2 equations partly consist of the same terms, but also of specific Level-2 equation terms:

- `B`*k* = beta component k, corresponding with the Level-1 regression coefficient. At this level, however, `B`*k* is an outcome variable.

- `G`*k* = gamma component k (γ_k). These are the fixed parameters to be estimated in the multilevel model.

- `V`*i* = one of the variables from the data file (as explained above). In this case, it is a Level-2 predictor variable. It means that this variable is considered to have the same value for all Level-1 units within a particular Level-2 unit. To be certain that this is the case, for each Level-2 variable the average is computed over all Level-1 units within the particular Level-2 unit. Note that this feature may be used to create an aggregated Level-1 variable, serving as a Level-2 predictor variable, simply by specifying a Level-1 variable as a Level-2 variable as well.

- `U`*k* = Level-2 random term k, i.e., u_{jk}, the k-th element of a typical u_j. As with the first level, this component is considered a residual or error term, but now for the second level. The second level may have more than one error term: one for each Level-2 equation (i.e., for each β element). The variances and the covariances of these terms have to be estimated from the data.

In the equations each term is followed by a number (except for the Level-1 random term `E`). For the `V`*i* term this number is the variable number, the position of the variable in the data file (e. g., `V4`, the fourth variable in the data file). The other terms only use a number for identification, without any additional meaning (e. g., `G3`, one of the fixed parameters). The `B`*k* terms have meaning in the equations of both levels. Every equation consists of one term before and at least one term after the equals sign.

Example:

```
/MODEL
  B1 = G1 + G2*V6 + U1  % random intercepts, dependent on level-2 predictor
  B2 = G3 + G4*V6 + U2  % random slopes, dependent on the same level-2 predictor
  V4 = B1 + B2*V5 + E   % level-1 equation, dependent on level-1 predictor
```

As shown above, terms on the right hand side of the equations are connected by plus signs. A variable and a corresponding parameter are connected by an asterisk (`*`). This is used to connect a fixed parameter and an observed predictor variable in Level-2 equations and to connect a Level-1 regression coefficient and an observed predictor variable in the Level-1 equation. In chapter 4, several variations of the two-level model will be presented and discussed in more detail.

Because a Level-1 equation and at least one Level-2 equation are required, the minimal specification of a model is:

```
/MODEL
  B1 = G1      % fixed intercept
  V4 = B1 + E  % level-1 variation
```

or

```
/MODEL
  B1 = U1      % random intercept
  V4 = B1 + E  % level-1 variation
```

3.4 `/CONSTRAINTS` (optional)

`MLA` has a limited option for imposing parameter constraints. The `/CONSTRAINTS` statement allows parameters to be fixed to a certain value. Constraints are imposed as: “parameter = value”. The parameter is held fixed during estimation and is not used for estimation of the standard errors either. The standard error will be zero and no *t*-test is performed for this parameter. This feature is still only implemented for the FIML estimation part with the BFGS estimation method. It is simply ignored for the various OLS estimators, but when requesting constraints with the EM estimation method, the program stops with a fatal error.

Values must be specified as floating point numbers. Variances and covariances are specified by connecting the appropriate Level-2 residual terms by an asterisk.

Example:

```
/CONSTRAINTS
  G1    = 1.0  % fix component G1 to 1.0
  U1*U1 = 0.5  % fix level-2 variance of U1 to 0.5
  U1*U2 = 0.0  % fix level-2 covariance U1*U2 to 0.0
```

Requests for constraints on the elements of the Level-2 covariance matrix Θ are only handled well if Θ is not reparameterized (see the reparameterization subcommand of the /TECHNICAL command below). No error is given if both are requested, but the results are typically incorrect. This will be fixed in a future version of MLA.

3.5 /TECHNICAL (optional)

The /TECHNICAL statement provides useful possibilities to alter the estimation process. It concerns the estimation method (estimation), minimization algorithm (minimization), the reparameterization of the parameters to ensure positive definiteness of estimated covariance matrices (reparameterization), the maximum number of warnings (warnings), the maximum number of iterations (maxiter), the convergence criterion (convergence), the random seed to be used for the simulations (seed), and the possibility of writing intermediate iteration results to disk (file). If this statement and subsequent substatements are not specified, the program will run using default values.

Example:

```
/TECHNICAL
  estimation  = fiml      % estimation method fiml
  maxiter     = 10        % maximum number of iterations equals 10
  convergence = 0.00001   % function convergence set to 0.00001
  file        = tech.out  % technical results will be written to tech.out
```

3.5.1 estimation (optional)

The substatement estimation provides the opportunity to set the estimation method. One can choose between fiml and reml. The default method is fiml, which represents full information maximum likelihood estimation; reml is restricted maximum likelihood estimation. Both procedures are described in chapter 2.

3.5.2 minimization (optional)

This substatement sets the minimization method. One can choose between BFGS, using the Broyden-Fletcher-Goldfarb-Shanno variant of the quasi-Newton minimization method (e.g., Nocedal & Wright, 1999, chap. 8), and EM, the Expectation-Maximization algorithm (Dempster, Laird, & Rubin, 1977). The default minimization method is BFGS, which tends to be both fast and stable. However, for REML estimation, only EM is currently supported.

3.5.3 `reparameterization` (optional)

The Level-2 covariance matrix should be a positive (semi-)definite matrix. To impose this restriction, the parameters can be written in the following way: $\Theta = LDL'$, where Θ is the covariance matrix, L is a lower triangular matrix with ones on the diagonal, and D a diagonal matrix with nonnegative diagonal elements. `MLA` offers two reparameterizations with which the latter can be accomplished.

The "root" reparameterization writes $D_{kk} = \delta_k^2$. The parameters actually used in the minimization are the subdiagonal elements of L and the δ's, i.e., the square roots of the diagonal elements of D. This method is the default, and it can be requested explicitly by `reparameterization = root`.

The "logarithm" reparameterization writes $D_{kk} = \exp(\delta_k)$. Then, the parameters actually used in the minimization are the subdiagonal elements of L and the δ's, which are now defined as the (natural) logarithms of the diagonal elements of D. This method can be requested by `reparameterization = logarithm`.

In the output, however, the reparameterization is reversed, and the estimates of Θ are presented, with their corresponding standard errors.

By specifying `reparameterization = none`, no reparameterization is done. This has the drawback that the estimated covariance matrix may not be positive definite (and that the minimization algorithm may thus be less stable), but the advantage that constraints may be imposed on its elements, see the discussion of the `/CONSTRAINTS` command.

3.5.4 `warnings` (optional)

If the maximum number of warnings is reached, the program terminates execution. This substatement can change the default value of 25. The value must be an integer between 1 and 32767.

3.5.5 `maxiter` (optional)

The maximum number of iterations in the minimization process. The default value is 100. This number should be sufficient for reaching convergence if the sample size is large enough and/or the number of parameters to be estimated is not too large. Changing the minimization method or the convergence criterion (see below) can make it necessary to raise the maximum number of iterations. The value must be an integer between 1 and 32767.

3.5.6 `convergence` (optional)

After each iteration the new function value is compared to the previous function value. The obtained difference is compared to a `convergence` related value. If

$$\frac{|F_{i-1} - F_i|}{(|F_i| + |F_{i-1}|)/2} \leq \texttt{convergence},$$

convergence is said to have been reached. In this formula, F_i is the function value after the i-th iteration. The left-hand side of the inequality represents the ratio between the difference of two successive function values and the mean of these values. The default value of `convergence` is `1.0e-10`, i.e., 10^{-10}, and permitted values range from 0.0 to 1.0.

3.5.7 `seed` (optional)

For diagnostic purposes, one can provide an initial number (seed) for the random number generator. This is specified by the substatement `seed`. Using the same initial seed, the simulation samples will be identical. The seed value must be an integer between 1 and 1,073,735,823. The random number generator used is the Mersenne Twister with improved initialization (Matsumoto & Nishimura, 1998; Nishimura & Matsumoto, 2002).

3.5.8 `file` (optional)

The technical output can be written to a separate file. The file is specified after the `file` substatement under the `/TECHNICAL` statement and must be a valid file name. Only the essential information is written to this file. Its content changes over time and some inspection will show what is written in this file.

3.6 `/SIMULATION` (optional)

Several options for simulation are available in `MLA`. These are jackknife, bootstrap, and permutation. Theoretical details concerning the implementation of these resampling methods for the two-level model can be found in chapter 2.

With the substatements provided with the `/SIMULATION` statement, one can choose between the different kinds of simulation (using the keyword `kind`), and specify special simulation features (using the keywords `method`, `type` and `resample`). Additional features are the number of replications and the initial seed for the random number generator (`replications` and `seed`). Finally, one can specify a separate output file for intermediate results of the simulation (`file`).

Example:

```
/SIMULATION
  kind         = bootstrap  % use simulation method bootstrap
  method       = error      % resample from error vectors
  type         = raw        % use raw residuals as error vectors
  resample     = 1          % only resample level-1 units
  replications = 200        % repeat simulation 200 times
  seed         = 1041245    % start with random seed 1041245
  file         = boot.out   % write simulation results to boot.out
```

3.6.1 `kind` (required)

With this substatement the user can choose from three options, namely `bootstrap`, `jackknife`, and `permutation` simulation. All types of simulation work as follows:

- perform the analysis on the original data
- obtain a (new) sample
- repeat the analysis on the new sample
- save the (new) estimates

The last three steps, together called a replication, are repeated a number of times. Afterwards, bias-corrected estimates of model parameters and nonparametric estimates of standard errors are computed. These estimates are computed from the set of saved estimates and the original maximum likelihood estimates. Furthermore, for the bootstrap, confidence intervals are computed from the replications.

The bootstrap, the jackknife, and the permutation option differ in the way a new sample is obtained. The choice between bootstrap, jackknife, or permutation resampling also determines the way the final simulation estimates are computed. More details can be found in Chapter 2.

3.6.2 `method` (optional)

This substatement specifies the method of bootstrap to be performed. It is required whenever `kind = bootstrap`. One can choose between three different methods: `residuals`, `cases`, and `parametric`. The three methods differ in the way the bootstrap sample is obtained. They are described in more detail in chapter 2. A synonym for `residuals` is `error`, so in the following whenever `residuals` is mentioned, `error` is also implied.

`residuals` (or `error`)

This method resamples the elements of the (estimated) Level-1 and Level-2 error vectors. Subsequently a new outcome or dependent variable is computed using these error vectors, the original predictor or independent variables and their corresponding parameter estimates. With this method, the `type` substatement can be used to choose which type of estimated errors is used for resampling, see below.

`cases`

Using this method a bootstrap sample is created by resampling the original data. Thus, complete cases are randomly drawn (with replacement) from the original cases. The procedure follows the nested structure in the data, by a nested resampling of cases: Level-2 units are randomly drawn (with replacement) and cases within a particular drawn unit are resampled. It is also possible to resample only complete Level-2 units, where the Level-1 units within a sampled Level-2 units are the same as in the original data set (which is useful for repeated measures data), or to resample only Level-1 units within Level-2 units, where the Level-2 units are the same as in the original sample, but the Level-1 units within each Level-2 units are resampled (useful when there are few Level-2 units and many Level-1 units in each Level-2 unit, such in studies with many subjects from a few countries).

`parametric`

This method computes a new outcome or dependent variable using the original predictor variables, their corresponding parameter estimates and a set of random Level-1 and Level-2 error terms. These random terms are obtained as follows: New Level-1 errors are drawn from a normal distribution with mean zero and variance $\widehat{\sigma}^2$, which is the original estimate of the Level-1 variance component. New Level-2 errors are drawn from a (multivariate) normal distribution with zero mean vector and covariance matrix $\widehat{\Theta}$, which contains the original estimates of the Level-2 variance components.

3.6.3 type (optional)

The substatement type is only required whenever the substatement kind = bootstrap is used in combination with method = residuals. The type substatement specifies the type of estimation that is used to determine the Level-1 and Level-2 residuals. One can choose between raw, shrunken, bartlett, green, and mcdonald. More details can be found in Chapter 2. The way in which the Level-1 and Level-2 error terms are estimated from the "total" residuals is discussed in chapter 2.

3.6.4 balancing (optional)

For the bootstrap methods residuals and cases, a balanced bootstrap can be specified on this substatement. In that case balancing = balanced must be specified. Default is balancing = unbalanced.

3.6.5 resample (optional)

The substatement resample offers the user the choice at which level units will be resampled. The default is 0 for the bootstrap methods, which means that at both levels units will be resampled. If kind = bootstrap and method = cases, the user may choose 1 or 2, which means that only Level-1 units or only Level-2 units will be resampled, respectively. With kind = jackknife, there is no default and resample = 1 or resample = 2 must be chosen by the user.

Characteristics of the data gathering process and the data structure will determine which choice is appropriate. For instance, with repeated measures (Level-1) nested within individuals (Level-2), it is probably not useful to resample Level-1 units with the cases bootstrap. With multilevel data, the intraclass dependency will typically imply that the jackknife must be applied at Level-2.

3.6.6 linking (optional)

The Level-1 and Level-2 residuals can be drawn linked or unlinked during simulation. Linking the residuals means that the Level-1 residuals will be drawn from the same unit as where the Level-2 residual was drawn from. This is specified with linking = linked. Specifying linking = unlinked has the same result as not using the substatement at all. This is the default.

3.6.7 replications (optional)

Using the substatement replications the number of bootstrap replications is specified. It must be an integer value between 1 and 32767 ($2^{15} - 1$). The default value is 100. This number is usually considered sufficient for bias correction and computation of standard errors, but for computing bootstrap confidence intervals a value of 1000 or more is needed. This has also been concluded in another context by Markus (1994).

3.6.8 convergence (optional)

See the /TECHNICAL statement. Specifying the convergence substatement within the /SIMULATION statement has only implications for the convergence during simulation.

3.6.9 `file` (optional)

Results of the simulation analysis can be written to a file. Using the substatement `file`, a filename may be specified. Filenames must satisfy the ususal `DOS` conventions on filenames. For each replication, the following results are written to the file (in `ascii`, space separated):

1. global information
    - replication number
    - luxury level (obsolete)
    - seed
    - 0 (obsolete)
    - 0 (obsolete)
    - number of iterations until convergence
    - the minimum of the −2 log likelihood function
2. estimation results: triplets containing
    - parameter number
    - estimate
    - estimated variance of estimator (square of standard error)

   of each parameter. The parameters are in the following order: σ_ε^2, $\gamma_1, \ldots, \gamma_p$, Θ_{11}, Θ_{21}, Θ_{22}, $\Theta_{31}, \ldots, \Theta_{qq}$, where p is the dimension of γ and q is the dimension of each β_j.

The estimation results are thus repeated "`replications`" times. The results of the simulation analysis are used to compute the final bootstrap and jackknife estimates. The results of a replication are not taken into account when the algorithm did not converge or when the estimate or its standard error was fixed to zero because it reached the boundary of its parameter space. Further elaboration concerning this subject can be found both in the previous and in the next chapter.

3.7 `/INTERVAL` (optional)

Several options for bootstrap confidence interval estimation are available in `MLA`. Hence, this statement only has effect when the bootstrap is selected in the `/SIMULATION` command. The specific choices are made through a number of substatements.

3.7.1 `kind` (required)

With this substatement the user can choose from four methods, namely `normal` (bootstrap normal interval), `percentile` (Efron's percentile), `bias-corrected percentile`, and `bootstrap-t` (percentile-*t*).

3.7.2 `alpha` (optional)

This is α, i.e., 1 minus the confidence level of the confidence intervals. As usual, the confidence intervals are two-sided, with an estimated probability mass of $\frac{1}{2}\alpha \cdot 100\%$ on each side. The default value of `alpha` is 0.05, which gives 95% confidence intervals. Note that in some earlier versions of `MLA` `alpha` was half the value it is now (i.e., a 95% confidence interval was obtained with `alpha = 0.025`), but we think the current specification is more natural.

3.7.3 weight (optional)

This substatement has implications for the internal bootstrap, performed on the bootstrap-*t* confidence interval estimation. A balanced bootstrap can be specified on this substatement. In that case `weight = balanced` must be specified. Default is `weight = unbalanced`.

3.7.4 replications (optional)

As for the previous substatement, this substatement has also implications for the bootstrap-*t* method. The number of internal bootstrap replications is specified. It must be an integer value between 1 and 32767. The default value is 25.

3.7.5 convergence (optional)

See the `/TECHNICAL` statement. As for the previous two substatements, this substatement has only implications for the internal bootstrap in the bootstrap-*t* method.

3.7.6 file (optional)

Results of the interval estimation can be written to a file. Using the substatement `file`, a filename may be specified. Filenames must satisfy the usual `DOS` conventions on filenames.

3.8 /PRINT (optional)

The `/PRINT` statement gives the user control over the output. Not all output is optional. The default output consists of a title page, an echo of the input, the maximum likelihood estimates (`FIML`), and system information. Output for the simulation analysis is generated whenever the `/SIMULATION` statement is used. Additional output is controlled by substatements following the `/OUTPUT` statement. These will be briefly explained below. A more profound elaboration follows in chapter 4. Most theory underlying the different parts of the output can be found in chapter 2.

Example:

```
/PRINT
  input                        = yes        % display digested input statements
  descriptives                 = V1         % display variable descriptive statistics
  random level-1 coefficients = B1,sigma
  olsquares                    = no
  residuals                    = U2,E
  posterior means              = B1
  diagnostics                  = yes
```

3.8.1 input (optional)

The subcommand `input = yes` requests extra information about the input and the output. Specifically, the input statements are digested and re-displayed, and a single equation form of the model, as in (2.3), is displayed, and all used and unused options are spelled out, from which the input can be checked.

3.8.2 `descriptives` (optional)

If this is requested, a few simple summary statistics are displayed. After the keyword `descriptives` the user may specify both variables and Level-2 identification codes; `descriptives = all` means all variables and all Level-2 units.

For the total sample and for every Level-2 unit specified, the following statistics are computed and displayed: mean, standard deviation (denoted by `Stddev`), variance, skewness, kurtosis, Kolmogorov-Smirnov's Z (denoted by `K-S Z`), significance level of Z (denoted by `Prob(Z)`), minimum, 5th-percentile (`P5`), first quartile (`Q1`), median, third quartile (`Q3`), 95th percentile and the maximum. Formulas have been given in section 2.2.

3.8.3 `random level-1 coefficients` (optional)

The random Level-1 coefficients or Level-2 outcomes consist of ordinary least squares estimates per Level-2 unit. Estimates of the regression coefficients and estimates of the error variance, including their standard errors, t-ratios and exceedance probabilities of the t-ratios per Level-2 unit are displayed in separate blocks with their Level-2 unit number and Level-2 unit size. After the keyword B's and sigma may be specified.

3.8.4 `olsquares` (optional)

This part contains the ordinary least squares estimates for the fixed (`G`k) and random (variances and covariances of `U`k and `E`) parameters. A regression analysis is performed, ignoring grouping, to obtain the former. For the error variance two estimates are displayed, the one-step (`E(1)`) and two-step (`E(2)`) estimates, corresponding to (2.20) and (2.28), respectively. The estimate of the covariance matrix of `U` is obtained from (2.24).

3.8.5 `residuals` (optional)

After the keyword both Level-1 and Level-2 residuals may be specified (`U` and `E`). For the first level, three different types of residuals are displayed, namely the total, raw, and shrunken residuals. The Level-2 residuals are the raw and shrunken residuals for the specified Level-2 components. These estimates are based on the `FIML` or `REML` estimates. Formulas are given in section 2.7.

3.8.6 `posterior means` (optional)

Displayed are the posterior means (2.35) which are specified following the keyword. These estimates are based on the `FIML` or `REML` estimates of the parameters.

3.8.7 `diagnostics` (optional)

If `diagnostics = yes` is requested, several diagnostics are printed. First, it prints the sample sizes at both levels, the mean of the within-group sample sizes, and the "effective sample size". The latter is computed according to the formula ESS $= N/(1+(N/J-1)\rho)$, where $\rho = \Theta^*/(\Theta^*+\sigma_\varepsilon^2)$ is the conditional intraclass correlation and Θ^* is the diagonal element of Θ corresponding to the Level-1 intercept. See section 2.11 for a critical discussion about the usefulness of these statistics.

Second, it prints four experimental pseudo-R^2 measures. The properties of these are currently unknown. They are implemented for experimental reasons only. In a future version of `MLA`, they will either be removed or more extensively documented.

Finally, the Level-1 and Level-2 "outliers" are printed. See section 2.9 for the formulas of the statistics and a discussion of their properties. Actually, the term "outlier" is a little misleading here, because they are printed for all Level-1 and Level-2 units, whether they should be considered outliers or not.

3.9 `/PLOT` (optional)

The `/PLOT` statement gives the user control over some plot options. The output of `MLA` consists only of `ASCII` files, so these are rough plots, but they may be useful for quick diagnostic purposes. For high-quality graphics, the numerical output of `MLA` and its options for writing all kinds of results to files, can be invoked. This output can then be used as input to high-quality graphics software.

3.9.1 `histograms` (optional)

This option is only in effect when `/SIMULATION` is chosen. If so, all parameters may be specified and histograms will be displayed of the estimates (of the different replications) of the specified parameters.

3.9.2 `scatters` (optional)

Scatterplots can be obtained for prediction and residuals. Specifying `prediction` produces a scatterplot of the response variable versus the predicted values based on the estimated fixed parameters. Specifying a variable produces a scatterplot of this variables versus all residuals associated with this variable.

Chapter 4

Worked examples

In this chapter, a few examples will be discussed. In each example, the input file will be given and a relevant piece of the output file will be shown and discussed. The input files and corresponding data files are included in the MLA distribution, so these analyses can be repeated by the user. The output of MLA consists of a single text file, which is the second parameter in the statement that starts program execution, see p. 41. The output consists of several parts, and each part starts with a brief heading.

The analyses presented here range from a simple analysis of variance to a bootstrap analysis for a complicated two-level model. It is not our intention to give extensive examples of case studies. Rather, the examples discussed here are intended to give insight in how to use MLA for different analyses and glance at specific parts of the output.

4.1 Random effects analysis of variance

To illustrate how to run a random effects ANOVA using MLA, we consider part of the *Sesame Street* data set. The original data set from Glasnapp and Poggio (1985) is used in Stevens (1990) for an analysis of covariance. The original data set included 12 background variables and 8 achievement variables for 240 children from 5 sites. Here, we only use the data from the first 3 sites, and only consider the achievement variable measuring knowledge of numbers. This variable was measured on two occasions. In between these occasions, the children watched a series of the TV program Sesame Street. This series intended to teach pre-school skills to 3 to 5 year old children.

We will now perform an analysis of variance on these data. Here, Level-2 (j) indicates the site and Level-1 (i) the child. The model to be estimated is

$$y_{ij} = \gamma + u_j + \varepsilon_{ij}, \tag{4.1}$$

where γ is the overall mean on the posttest score, u_j is the Level-2 deviation from γ, or Level-2 error component, and ε_{ij} is the Level-1 deviation from $\gamma + u_j$, the average score of unit j, also called the Level-1 error component. Equation (4.1) can be divided into two separate equations, one for each level:

$$y_{ij} = \beta_j + \varepsilon_{ij},$$
$$\beta_j = \gamma + u_j.$$

In this way, the deviations or error components for the different levels are easily seen. These equations are also the equations that are to be used in MLA to specify the model. Along with the other statements, the input file is as follows:

```
/TITLE
  analysis of variance
/DATA
  file      = sesame.dat
  variables = 3
  id2       = 1
/MODEL
  b1 = g1 + u1
  v3 = b1 + e
/PRINT
  descriptives = all
  olsquares    = yes
/END
```

The top of the output is the `MLA` title page. It supplies information about the name and origin of the program, and cannot be suppressed.

```
             MMMM        MMMMM  LLLL  AAAAAAAA
            MMMMM      MMMMMM  LLLL  AAAAAAAAAA
           MMMM M    MMMMMMM  LLLL  AAAA    AAAA
          MMMM  MM MMM MMMM  LLLL  AAAA      AAAA
         MMMM   MMMM  MMMM  LLLL  AAAA        AAAA
        MMMM    MM   MMMM  LLLL  AAAAAAAAAAAAAAAAAA
       MMMM     M   MMMM  LLLL  AAAAAAAAAAAAAAAAAAAA
      MMMM         MMMM  LLLL  AAAA              AAAA
     MMMM         MMMM  LLLL  AAAA                AAAA
    MMMM         MMMM  LLLL                        AAAA
   MMMM         MMMM  LLLLLLLLLLLLLLLLLLLLLLLLLLLL  AAAA
  MMMM         MMMM  LLLLLLLLLLLLLLLLLLLLLLLLLLLLLL  AAAA
                                                      AAAA
  Multilevel Analysis for Two Level Data               AAAA
                                                        AAAA
  Version 4.1                                            AAAA
                                                          AAAA
  Developed by                                             AAAA
    Frank Busing                                            AAAA
    Erik Meijer                                              AAAA
    Rien van der Leeden                                       AAAA
                                                               AAAA
  Published by                                                  AAAA
    Leiden University                                            AAAA
    Faculty of Social and Behavioural Sciences                    AAAA
    Department of Psychometrics and Research Methodology           AAAA
    Wassenaarseweg 52                                               AAAA
    P.O. Box 9555                                                    AAAA
    2300 RB Leiden                                                    AAAA
    The Netherlands                                                    AAAA
    Phone +31 (0)71-5273761                                             AAAA
    Fax   +31 (0)71-5273619                                              AAAA
```

Except for the title page and the optional `input` part, every part contains a header. The header is always the same and is made of a few lines of standard text and the title of the analysis, as supplied by the user. For this first example, it reads:

```
  MLA (R)  Multilevel Analysis for Two Level Data  Version 4.1       12-15-2005
  Copyright 1993-2005 Leiden University  All Rights Reserved            Part  2

  Thu Dec 15 10:15:27 2005
  analysis of variance
```

The first proper part (Part 1) of the output contains an echo of the input file statements. This part is always included in an output file.

```
Inputfile statements

 1  /TITLE
 2    analysis of variance
 3  /DATA
 4    file      = sesame.dat
 5    variables = 3
 6    id2       = 1
 7  /MODEL
 8    b1 = g1 + u1
 9    v3 = b1 + e
10  /PRINT
11    input        = yes
12    descriptives = all
13    olsquares    = yes
14  /END

    14 lines read from "anova.in"
```

Part 2 is the first optional part of the output. It is triggered by the `input` keyword under the `/PRINT` statement. It contains extra information about the input and the output. Specifically, the input statements are digested and re-displayed and all used and unused options are spelled out.

```
Input information

  Required

    Name of datafile       : SESAME.DAT
    Number of variables    : 3
    Level-2 id. column     : 1
    Equation  1            : B1=G1+U1
    Equation  2            : V3=B1+E
    Single equation        : V3=E0+G1+U1

  Optional

    Title of analysis      : analysis of variance
    Level-1 id. column     : 0
    Centering              :
    Level-2 centering      :
    Estimation method      : 1
    Minimization method    : 1
    Reparameterization     : 1
    Maximum iterations     : 100
    Convergence            : 1e-010
    Warnings (maximum)     : 25
    Kind of simulation     : 0
    Simulation method      : 0
    Simulation balance     : 0
    Simulation linking     : 0
    Residuals type         : 0
    Resampling type        : 0
    Initial random seed    : 0
    Simulation convergence : 1e-010
    Number of replications : 0
    Simulation output file :
    Kind of CI estimation  : 0
    CI alpha               : 0.025000
    CI convergence         : 1e-010
    CI replications        : 25
    Print input            : 1
    Print explore          : 1, ALL
    Print olsq             : 1
    Print outcomes         :
    Compute residuals      : 0
    Print residuals        :
```

```
Print posterior means  :
Print diagnostics      : 0
Print intervals        : 0
Max equations          : 2
Level-1 size           : 1
Level-2                : 1
X-size                 : 1
Z-size                 : 1
Parameters             : 3
Level-2 parameters     : 1
Input file             : anova.in
Output file            : anova.out
Verbose                : 0
Monte Carlo            : 0
Monte Carlo file       :
Plot histograms        :
Plot scatterplots      :
Response variable      : 3
Explanatory variables  :  0(1)
Random level-2 vars.   :  0(1)
Random level-1 coeffs. :  0(1)
Level-2 outcome  1     :  0(1)[1]
```

The single equation shows the integration of the Level-2 equations and the Level-1 equation in the same way as in chapter 2, equations (2.1), (2.2), and (2.4). It is displayed directly below the model specification.

Part 3 consists of the data descriptives, optionally given by the use of the keyword `descriptives = all` under the `/PRINT` statement. These statistics are displayed in two major blocks, and are preceded by the number of Level-1 and Level-2 units.

```
Data descriptives

Data descriptives for all units
# Level-1 units         = 179
# missing Level-1 units = 0
# correct Level-1 units = 179
# correct Level-2 units = 3

Var      Mean    Stddev  Variance  Skewness  Kurtosis     K-S Z   Prob(Z)

  1      2.02      0.83      0.70     -0.04     -1.57      3.18      0.00
  2     21.37     10.92    119.25      0.72     -0.16      1.45      0.03
  3     31.02     12.89    166.19     -0.07     -1.13      1.37      0.05

Var   Minimum        P5        Q1    Median        Q3       P95   Maximum

  1      1.00      3.00      3.00      3.00      3.00      3.00      3.00
  2      4.00     52.00     52.00     52.00     52.00     52.00     52.00
  3      0.00     54.00     54.00     54.00     54.00     54.00     54.00

Data descriptives for level-2 unit 1
# Level-1 units = 60

Var      Mean    Stddev  Variance  Skewness  Kurtosis     K-S Z   Prob(Z)

  1      1.00      0.00      0.00      0.00      0.00      3.87      0.00
  2     22.22     13.26    175.90      0.60     -0.85      1.28      0.08
  3     30.08     13.93    194.04      0.09     -1.38      1.13      0.16

Var   Minimum        P5        Q1    Median        Q3       P95   Maximum

  1      1.00      1.00      1.00      1.00      1.00      1.00      1.00
  2      4.00     52.00     52.00     52.00     52.00     52.00     52.00
  3      6.00     54.00     54.00     54.00     54.00     54.00     54.00
```

```
Data descriptives for level-2 unit 2
# Level-1 units = 55

Var       Mean    Stddev  Variance  Skewness  Kurtosis     K-S Z   Prob(Z)

  1       2.00      0.00      0.00      0.00      0.00      3.71      0.00
  2      26.04      9.50     90.26      0.51     -0.61      0.97      0.30
  3      38.53     10.65    113.48     -0.77     -0.15      0.89      0.41

Var    Minimum        P5        Q1    Median        Q3       P95   Maximum

  1       2.00      2.00      2.00      2.00      2.00      2.00      2.00
  2      11.00     48.00     48.00     48.00     48.00     48.00     48.00
  3       8.00     53.00     53.00     53.00     53.00     53.00     53.00

Data descriptives for level-2 unit 3
# Level-1 units = 64

Var       Mean    Stddev  Variance  Skewness  Kurtosis     K-S Z   Prob(Z)

  1       3.00      0.00      0.00      0.00      0.00      4.00      0.00
  2      16.56      7.19     51.65      0.55     -0.10      1.05      0.22
  3      25.44     10.43    108.85      0.12     -0.78      0.79      0.56

Var    Minimum        P5        Q1    Median        Q3       P95   Maximum

  1       3.00      3.00      3.00      3.00      3.00      3.00      3.00
  2       4.00     37.00     37.00     37.00     37.00     37.00     37.00
  3       0.00     46.00     46.00     46.00     46.00     46.00     46.00
```

The first variable is the Level-2 identifier variable. The second and third variables are the score on the pretest and the posttest, respectivily. Formulas defining these descriptives can be found in section 2.2.

Part 4 gives the optional OLS estimates, which are requested by the `olsquares = yes` option in the /PRINT statement. As described in section 2.4, ordinary least squares estimation yields two different estimates for the Level-1 variance component, σ^2, one by ignoring the hierarchical data structure and one using this structure. These are both displayed in this part of the output, with the one-step estimate labeled E(1) and the two-step estimate labeled E(2). U1*U1 gives the variance estimate for the Level-2 variance component U1.

```
Ordinary least squares estimates

Fixed parameters

   Label       Estimate             SE

      G1      31.016760       0.963540

Random parameters

   Label       Estimate             SE

    E(1)     166.185111      17.615587

   U1*U1      29.469076      24.061400

    E(2)     136.503030      14.469292


E(1): one-step estimate of sigma squared (ignoring grouping)
E(2): two-step estimate of sigma squared
See documentation for further elaboration on these subjects
```

As can be seen, the overall mean (G1) equals the mean of Variable 3, the score on the posttest (31.02). Ignoring grouping will result in 166.19 for σ^2. Using the two-step procedure lowers the estimate to 136.50 and also gives an estimate of the variance of u_j (29.47).

Part 5 contains the FIML estimates. This part is default and appears in all output. Compared to the previous ordinary least squares estimates part, T-values and probabilities for T are given. Here, unlike for the OLS estimates, these are theoretically justified.

```
Full information maximum likelihood estimates (BFGS)

Fixed parameters

   Label       Estimate            SE             T        Prob(T)

      G1      31.322474      3.123584         10.03         0.0000

Random parameters

   Label       Estimate            SE             T        Prob(T)

   U1*U1      26.935248     23.900119          1.13         0.2597

       E     138.833328     14.799679          9.38         0.0000


Intra-class correlation = 26.94/(138.83+26.94) = 0.1625

# iterations = 5
-2*Log(L)    = 1398.626571
```

Note that the estimates are very close to the two-step OLS estimates. Whenever there are residuals associated with the grand mean, correlation is computed and given just below the FIML estimates. The formula for the intra-class correlation is

$$\rho = \frac{\tau}{\tau + \sigma^2}$$

and in MLA notation,

$$\rho = \frac{\texttt{U1*U1}}{\texttt{U1*U1 + E}}.$$

If the technical keyword is omitted from the /PRINT statement a short description of the final iteration results is given in the FIML part. Here, convergence is reached in 5 iterations and yields a -2*Log(L) value of 1398.63.

The final part of the output contains some information about whether everything went well. This can also not be suppressed. As we can see here, the program is terminated correctly in less than 0.01 seconds.

```
This job required lots of memory
and took 0.00 seconds of processor time

   0 warning(s) issued
   0 error(s) detected

End of job.
```

4.2 Random effects analysis of covariance

For the next example the same Sesame Street data set is used, but now a random effects analysis of covariance is performed on these data. The model to be estimated is

$$y_{ij} = \gamma_1 + \gamma_2 X_{ij} + u_j + \varepsilon_{ij}, \tag{4.2}$$

where γ_1 is the intercept, X_{ij} is the covariate with slope γ_2, u_j is the Level-2 error component and ε_{ij} is the Level-1 error component. Equation (4.2) can be divided into separate equations, one equation for Level-1 and in this case two Level-2 equations:

$$\begin{aligned} y_{ij} &= \beta_{1j} + \beta_{2j} X_{ij} + \varepsilon_{ij}, \\ \beta_{1j} &= \gamma_1 + u_j, \\ \beta_{2j} &= \gamma_2. \end{aligned}$$

Along with the other statements, the input file now becomes:

```
/TITLE
  analysis of covariance
/DATA
  file      = sesame.dat
  variables = 3
  id2       = 1
/MODEL
  b1 = g1 + u1
  b2 = g2
  v3 = b1 + b2*v2 + e
/PRINT
  olsquares = yes
/END
```

Note that in the equation where `b2` is the "outcome", there is no error term. This ensures that `b2` is a fixed coefficient. Here is the output of the `OLS` estimation part:

```
  Ordinary least squares estimates

  Fixed parameters

     Label        Estimate             SE

        G1       14.672451       1.621040
        G2        0.764871       0.067590

  Random parameters

     Label        Estimate             SE

      E(1)       96.968087      10.307591

     U1*U1        7.217027       5.892678

      E(2)       88.980063       9.458474


  E(1): one-step estimate of sigma squared (ignoring grouping)
  E(2): two-step estimate of sigma squared
  See documentation for further elaboration on these subjects
```

Compared to the previous example, a fixed parameter (`G2`) is added in the `OLS`-estimates part. This is the regression coefficient of the Level-1 covariate containing the pretest score.

The parameter estimate for the regression coefficient of the covariate is also added to the FIML output part. The additional T-value and Prob(T) indicate that the pretest variable explains a significant part of the variance of the posttest variable (T = 10.18, Prob(T) = 0.0000).

```
Full information maximum likelihood estimates (BFGS)

Fixed parameters

   Label       Estimate              SE               T         Prob(T)

      G1      16.196937        2.226470            7.27          0.0000
      G2       0.699891        0.068761           10.18          0.0000

Random parameters

   Label       Estimate              SE               T         Prob(T)

   U1*U1       6.766703        6.759617            1.00          0.3168

       E      89.831170        9.576024            9.38          0.0000


Intra-class correlation = 6.77/(89.83+6.77) = 0.0701

# iterations = 7
-2*Log(L)    = 1318.217264
```

Entering the covariate into the analysis is justified, because it has a statistically significant non-zero effect. The same justification could be made with the use of the likelihood-ratio test. This test is based on the fact that the difference between minus two times the loglikelihood function value (-2*Log(L)) of two nested models follows a chi-square distribution with the number of degrees of freedom equal to the difference in the number of free parameters. The two models (example 1 and example 2) are nested and the likelihood-ratio test can be applied. The difference between the function values is approximately 1399 – 1318 = 81, and the degrees of freedom is equal to 1. The likelihood-ratio test corroborates that the effect is highly significant.

4.3 Repeated measures analysis

Multilevel analysis can often be an appropriate analysis method for repeated measures data see, e.g., Van der Leeden (1998). Here, we illustrate the use of MLA for repeated measures data on the frequently analyzed *Rat* data set. The first use of these data with multilevel analysis appeared in (Strenio, Weisberg, & Bryk, 1983).

The Rat data consist of the weights of 10 rats. These rats were measured five times with four week intervals from birth. Also included in the model is the weight of each rat's mother (V2). Divided into two levels, the equations are given by

$$
\begin{aligned}
y_{ij} &= \beta_{1j} + \beta_{2j} X_{ij} + \varepsilon_{ij}, \\
\beta_{1j} &= \gamma_1 + \gamma_2 W_j + u_{1j}, \\
\beta_{2j} &= \gamma_3 + \gamma_4 W_j + u_{2j},
\end{aligned}
$$

where X_{ij} (V5) is the age (in weeks, divided by 4, minus 2, so that it is in deviation of the mean) of the rat, and W_j (V2) represents the weight of the mother. The input file for the repeated measures example is as follows.

```
/TITLE
  Rat Data - Repeated Measures - 5 timepoints
/DATA
  file      = rat.dat
  variables = 4
  id2       = 3
  center    = v4
/MODEL
  b1 = g1 + g2*v2 + u1
  b2 = g3 + g4*v2 + u2
  v1 = b1 + b2*v4 + e
/TECH
  estimation   = reml
  minimization = em
  maxiter      = 512
/PRINT
  olsquares             = yes
  random level-1 coeffs = all
  residuals             = u1,u2,e
  posterior means       = all
  diagnostics           = yes
/END
```

For each rat a multiple regression analysis is performed and displayed in the Random Level-1 coefficients part. This part is optional and displayed through the use of the `random level-1 coefficients` keyword in the `/PRINT` statement.

```
  Random Level-1 coefficients: ordinary least squares estimates per level-2 unit

  Parameter B1

  Unit    Size       Estimate            SE             T        Prob(T)

     1       5       111.4000        4.9044         22.71         0.0000
     2       5       120.2000        2.9967         40.11         0.0000
     3       5       119.8000        6.7621         17.72         0.0000
     4       5       103.4000        3.8018         27.20         0.0000
     5       5       100.0000        3.2701         30.58         0.0000
     6       5        99.0000        4.4505         22.24         0.0000
     7       5        93.0000        5.5281         16.82         0.0000
     8       5       113.6000        1.6391         69.31         0.0000
     9       5        90.4000        4.5284         19.96         0.0000
    10       5       121.0000        2.4549         49.29         0.0000

  Mean               107.1800
  Variance           132.6884

  Parameter B2

  Unit    Size       Estimate            SE             T        Prob(T)

     1       5        28.8000        3.4679          8.30         0.0000
     2       5        28.1000        2.1190         13.26         0.0000
     3       5        36.3000        4.7816          7.59         0.0000
     4       5        27.2000        2.6882         10.12         0.0000
     5       5        23.4000        2.3123         10.12         0.0000
     6       5        29.3000        3.1470          9.31         0.0000
     7       5        25.6000        3.9090          6.55         0.0000
     8       5        19.7000        1.1590         17.00         0.0000
     9       5        23.6000        3.2021          7.37         0.0000
    10       5        25.6000        1.7359         14.75         0.0000

  Mean                26.7600
  Variance            19.7138

  Parameter SIGMA
```

```
Unit    Size       Estimate             SE              T        Prob(T)

   1       5       120.2667        98.1973           1.22         0.2207
   2       5        44.9000        36.6607           1.22         0.2207
   3       5       228.6333       186.6783           1.22         0.2207
   4       5        72.2667        59.0055           1.22         0.2207
   5       5        53.4667        43.6554           1.22         0.2207
   6       5        99.0333        80.8604           1.22         0.2207
   7       5       152.8000       124.7607           1.22         0.2207
   8       5        13.4333        10.9683           1.22         0.2207
   9       5       102.5333        83.7181           1.22         0.2207
  10       5        30.1333        24.6038           1.22         0.2207

Mean                91.7467
Variance            64.4782

Note: random level-1 coefficients are also referred to as level-2 outcomes
See documentation for further elaboration on this subject
```

In the next part we can see that both `G2` and `G4` indicate that the mother's weight has a positive effect on the rat's weight. The rat's weight starts higher and rises faster with a heavier mother.

```
Restricted maximum likelihood estimates (EM)

Fixed parameters

   Label       Estimate             SE              T        Prob(T)

      G1      18.873660      21.002157           0.90         0.3688
      G2       0.545101       0.128963           4.23         0.0000
      G3       2.967709      11.848158           0.25         0.8022
      G4       0.146866       0.072753           2.02         0.0435

Random parameters

   Label       Estimate             SE              T        Prob(T)

   U1*U1      27.819366      21.183863           1.31         0.1891
   U2*U1      -8.248183       8.639480          -0.95         0.3397
   U2*U2       5.518709       6.985030           0.79         0.4295

       E      91.746606      23.688872           3.87         0.0001


Conditional intra-class correlation = 27.82/(91.75+27.82) = 0.2327

# iterations = 315
-2*Log(L)    = 384.359839
```

The formula used for the "conditional intra-class correlation" is still

$$\rho = \frac{\tau}{\tau + \sigma^2}$$

and in `MLA` notation,

$$\rho = \frac{\texttt{U1*U1}}{\texttt{U1*U1 + E}}$$

It is the intra-class correlation for observations if all Level-1 variables except the constant are equal to zero. In this case, given the centering of the age variable, this means the intra-class correlation in the middle age period, which may be an interesting value. See section 2.11 for a further discussion of the conditional intra-class correlation.

Note that the EM algorithm needs a huge number of iterations, whereas in the previous examples, the BFGS algorithm converged after only a few iterations. This slow convergence is a general characteristic of the EM algorithm and one of the reasons why MLA uses BFGS by default.

The posterior means may be compared with the Level-2 outcomes. As can be seen, the posterior means tend to be shrunken towards the grand mean, and therefore have less variance than the Level-2 outcomes.

```
Posterior means

Parameter B1

Unit       Estimate

   1       111.2477
   2       122.9870
   3       118.7899
   4       103.2974
   5       102.0644
   6        98.6570
   7        92.8984
   8       110.2970
   9        94.9923
  10       116.5688

Mean       107.1800
Variance   107.2725

Parameter B2

Unit       Estimate

   1        28.2171
   2        30.9801
   3        32.3522
   4        26.3462
   5        25.4355
   6        26.5173
   7        24.0949
   8        22.4476
   9        25.6699
  10        25.5391

Mean        26.7600
Variance     9.0631


Note: posterior means = shrunken estimates of random level-1 coefficients
See documentation for further elaboration on this subject
```

4.4 Multilevel analysis with bootstrap

Now, we discuss a complex input file for a two-level analysis with bootstrap resampling. The data file contains computer-generated data. It contains 231 Level-1 observations in total, in 15 Level-2 units. There are six variables: a constant, a Level-2 identifier, a Level-1 identifier, a Level-1 dependent variable, a Level-1 explanatory variable, and a Level-2 explanatory variable.

The model equations are given by

$$\begin{aligned}
y_{ij} &= \beta_{1j} + \beta_{2j} X_{ij} + \varepsilon_{ij}, \\
\beta_{1j} &= \gamma_1 + \gamma_2 W_j + u_{1j}, \\
\beta_{2j} &= \gamma_3 + \gamma_4 W_j + u_{2j},
\end{aligned}$$

where y_{ij} is the dependent variable (V4), X_{ij} (V5) is the Level-1 explanatory variable, and W_j (V6) is the Level-2 explanatory variable. The following input file shows the application of the bootstrap option of MLA, as well as several other options.

```
/TITLE
  Complete multilevel model
/DATA
  file      = multi.dat
  variables = 6
  id1       = 3
  id2       = 2
  missing   = v4(-0.6888)
  center    = v6
/MODEL
  b1 = g1    \\
     + g2*v6 \\
     + u1
  b2 = g3 + g4*v6 + u2
  v4 = b1 + b2*v5 + e
/TECHNICAL
  estimation         = fiml
  minimization       = bfgs
  reparameterization = root
  warnings           = 50
  maxiter            = 500
  seed               = 1041245
  convergence        = 1.0E-12
/SIMULATION
  kind               = bootstrap
  method             = residuals
  type               = shrunken
  balance            = unbalanced
  linking            = unlinked
  replications       = 200
/INTERVAL
  kind               = bias-corrected
  alpha              = 0.05
/PRINT
  input                 = yes
  descriptives          = v4,v5,v6,2,3
  olsquares             = yes
  random level-1 coeffs = all
  residuals             = u1,u2
  posterior means       = all
  diagnostics           = yes
/PLOT
  histograms = g2,g3,g4
  scatters   = predicted,v6,v5
/END
```

Note the declaration of the missing value on the dependent variable. This removes one observation from the analysis.

The following part gives the FIML estimates of the parameters.

```
  Full information maximum likelihood estimates (BFGS)
```

```
Fixed parameters

   Label       Estimate            SE             T        Prob(T)

      G1       0.936989      0.190592          4.92         0.0000
      G2       2.058120      0.327704          6.28         0.0000
      G3       0.956199      0.127996          7.47         0.0000
      G4       0.737448      0.219790          3.36         0.0008

Random parameters

   Label       Estimate            SE             T        Prob(T)

   U1*U1       0.467356      0.198821          2.35         0.0187
   U2*U1       0.050104      0.094816          0.53         0.5972
   U2*U2       0.160850      0.088703          1.81         0.0698

       E       1.096541      0.109506         10.01         0.0000


Conditional intra-class correlation = 0.47/(1.10+0.47) = 0.2988

# iterations = 10
-2*Log(L)    = 720.520970
```

An extract of the diagnostics is given below. Note the reduction of the sample size due to the missing value.

```
Diagnostics

Level-2 sample size      = 15
Total sample size        = 230
Mean Level-1 sample size = 15
Effective sample size    = 44

Squared correlation coefficients

Norm based R-squared         = 0.627456
Grand mean based R-squared   = 0.614517
Context mean based R-squared = 0.240201
Trimmed mean based R-squared = 0.485713

Level-1 outliers (sorted by Prob)

Level-1  Level-2  Level-1
   Unit     Unit     Unit             T          Prob

     86        6        5     -0.013931      0.988885
    160       11        9      0.010492      0.991629
     45        3        9      0.009729      0.992238
    ...      ...      ...         ...           ...
     26        2        6     -0.000077      0.999938
    155       11        4     -0.000071      0.999943
    147       10       10     -0.000013      0.999990

Level-2 Mahalanobis distances (sorted by Prob(M))

Unit             M        Prob(M)

   7      5.052856       0.080910
   4      3.803597       0.149863
  15      2.866620       0.238823
  11      2.709150       0.258324
   2      1.940617       0.379085
  12      1.340144       0.511717
   6      1.095992       0.578134
  10       1.062185       0.587987
   5       0.805649       0.668441
   3       0.760907       0.683561
   8       0.425112       0.808517
   1       0.410341       0.814510
  13       0.163562       0.921474
  14       0.075244       0.963077
   9       0.031498       0.984375


Effective sample size: N/(1+(N/J-1)*intra-class correlation)
Squared correlation coefficients (R-squared) are highly speculative in nature
Prob(M): probability - area under the curve of the chi-square distribution
See documentation for further elaboration on this subject
```

See section 2.11 for a discussion of the effective sample size.

The bootstrap estimates as printed as follows

```
Bootstrap estimates (unbalanced unlinked shrunken residuals)

  Replications done = 200
  Replications used = 200

Fixed parameters

   Label       Estimate             SE

      G1       0.920206       0.170909
      G2       2.037206       0.312473
      G3       0.958121       0.118839
      G4       0.742056       0.196784

Random parameters

   Label       Estimate             SE

   U1*U1       0.611998       0.144656
   U2*U1       0.060488       0.078446
   U2*U2       0.236573       0.053698

       E       1.208053       0.077424
```

The bootstrap confidence interval estimates are as follows

```
Confidence interval estimates (bias-corrected percentile method)

Fixed parameters

   Label       Estimate           Mean          Lower          Upper

      G1       0.936989       0.953771       0.648489       1.213440
      G2       2.058120       2.079033       1.553716       2.602604
      G3       0.956199       0.954277       0.780768       1.163196
      G4       0.737448       0.732839       0.406751       1.072559

Random parameters

   Label       Estimate           Mean          Lower          Upper

   U1*U1       0.467356       0.322714       0.372347       0.724392
   U2*U1       0.050104       0.039720      -0.060996       0.191503
   U2*U2       0.160850       0.085127       0.136783       0.253574

       E       1.096541       0.985029       1.075827       1.233065

Note: mean refers to average over bootstrap replications
See documentation for further elaboration on this subject
```

```
Bootstrap estimates (unbalanced unlinked shrunken residuals)

  Replications done = 200
  Replications used = 200

Fixed parameters

   Label       Estimate             SE       Confidence Interval

      G1       0.920206       0.170909     [ 0.648489   ,   1.213440]
      G2       2.037206       0.312473     [ 1.553716   ,   2.602604]
      G3       0.958121       0.118839     [ 0.780768   ,   1.163196]
      G4       0.742056       0.196784     [ 0.406751   ,   1.072559]

Random parameters

   Label       Estimate             SE       Confidence Interval

   U1*U1       0.611998       0.144656     [ 0.372347   ,   0.724392]
   U2*U1       0.060488       0.078446     [-0.060996   ,   0.191503]
   U2*U2       0.236573       0.053698     [ 0.136783   ,   0.253574]

       E       1.208053       0.077424     [ 1.075827   ,   1.233065]

Note: confidence intervals computed by bias-corrected percentile method
```

Figure 4.1 gives an example of the Level-1 scatterplots that are output, figure 4.2 gives an example of the Level-1 scatterplots that are output, and figure 4.3 gives an example of the histograms that are output. These are not state of the art high resolution graphics, but they may be useful for diagnostic purposes.

```
Scatterplots

    6.5941 +----------------+------------------------------------------+
           |                                                          1|
           |                                                1          |
           |                                                           |
           |                                                           |
           |                                     1      1              |
         p |                                              1            |
         r |                           1       1 1       1             |
         e |                     1     11  1    1       1              |
         d |                          1 2     1  21   1     1          |
         i |               1        11    1121 1   1                   |
         c |                    1 21   213 2 21   1                    |
         t |                 1      13 231 2 11 1   1                  |
         e |        1 1   2213432 122 11 1  2  1                       |
         d |          1 13  3131151 11  31   1                         |
           |      1 1  112111 12 1223  11111 1                         |
           |      113  4211231213 1   2                                |
           +11 1 1     3 312*2 122                                     +
           |      1   1 12112 111     1                                |
           |     1          1   1                                      |
           |  21 11122 2111                                            |
   -1.9893 +----------------+------------------------------------------+
         -3.5414                     observed                         8.5943

           Scatterplot predicted vs observed


    2.2941 +---------------------------+-------------------------------+
           |                                  1                        |
           |                      1                      1             |
           |                 1    2                                    |
           |    1           11           1   12  1 1      1    1       |
           |1  1   11  11   1  1   1  2  1  11 1  11         1         |
           |  1      2      1 1        2 1 1    21      1          1 1 |
         r |            1   21  2  1    1       2 121    111   1 1     |
         e |              1  1 1 311112  1 311  1  1 1     2 1         |
         s +      1     1 132 112  1   * 1 113  1   1       1        1 +
         i | 1         11  1  111    11 1212 113            1 1        |
         d |            113   2 11 113 21131 1122 1  1    1  1         |
         u |          1     1  2        1 1   1  11  1 1     1         |
         a |       1       1     2   11  1  1  11 2  1     1    1      |
         l |          1  11 1 1     1 11       12  1                   |
           |          1           1 1     21       1    1     1        |
           |   1       11    1  1             1                        |
           |                           1          1                    |
           |                                      1                    |
           |                                                           |
           |                                    1                      |
   -3.2004 +---------------------------+-------------------------------+
         -2.3101                    predictor                         2.5676

           Scatterplot predictor V5 vs residual E
```

Figure 4.1: An example of Level-1 scatterplots.

```
Scatterplots (resumed)

    1.0222 +---------------------------------+-------------------------+
           |                                                   1       |
           |                                                           |
           |                                                           |
           |                                      1                    |
           |                       1                                   |
           |                                                           |
         r |                                     1  1                  |
         e |               1     1                                     |
         s +1                             1  *                       1 +
         i |                           1                               |
         d |                                                           |
         u |                                                           |
         a |                                          1                |
         l |                                    1                      |
           |                                                           |
           |                                                           |
           |                                                           |
           |                            1                              |
           |                                                           |
           |                                         1                 |
   -1.4021 +---------------------------------+-------------------------+
         -1.3677                    predictor                         1.0436

           Scatterplot predictor V6 vs residual U1


    0.6003 +---------------------------------+-------------------------+
           |                                      1                    |
           |                                                           |
           |                                                           |
           |                                        1                  |
           |                                                           |
           |                     1                                     |
         r |                                                         1 |
         e |                                                           |
         s |                       1                                   |
         i |                                    1                      |
         d |1                                                          |
         u +                            1    *        1                +
         a |                              1                            |
         l |                           1                               |
           |                                                           |
           |                                                           |
           |                                                           |
           |               1                                   1       |
           |                                                           |
           |                                     1   1                 |
   -0.5131 +---------------------------------+-------------------------+
         -1.3677                    predictor                         1.0436

           Scatterplot predictor V6 vs residual U2
```

Figure 4.2: An example of Level-2 scatterplots.

```
Histograms

Count  Midpoint
    3   Extreme | (1.2472) (1.4059) (1.4347)
    5    1.5925 |xxxxx
    7    1.6784 |xxxxxxx
    6    1.7644 |xxxxxx
   20    1.8503 |xxxxxxxxxxxxxxxxxxxx
   17    1.9363 |xxxxxxxxxxxxxxxxx
   22    2.0223 |xxxxxxxxxxxxxxxxxxxxxx
   20    2.1082 |xxxxxxxxxxxxxxxxxxxx
   22    2.1942 |xxxxxxxxxxxxxxxxxxxxxx
   18    2.2801 |xxxxxxxxxxxxxxxxxx
   17    2.3661 |xxxxxxxxxxxxxxxxx
   11    2.4520 |xxxxxxxxxxx
   11    2.5380 |xxxxxxxxxxx
    8    2.6240 |xxxxxxxx
    5    2.7099 |xxxxx
    4    2.7959 |xxxx
    3   Extreme | (2.8032) (2.8839) (3.0018)
-----           +---------+---------+---------+---------+---------+
  200           0         10        20        30        40        50

                Histogram estimate G2


Count  Midpoint
    3   Extreme | (0.6904) (0.7041) (0.7061)
    6    0.7648 |xxxxxx
    3    0.7971 |xxx
   10    0.8294 |xxxxxxxxxx
   17    0.8617 |xxxxxxxxxxxxxxxxx
   16    0.8940 |xxxxxxxxxxxxxxxx
   20    0.9263 |xxxxxxxxxxxxxxxxxxxx
   16    0.9586 |xxxxxxxxxxxxxxxx
   21    0.9909 |xxxxxxxxxxxxxxxxxxxxx
   24    1.0232 |xxxxxxxxxxxxxxxxxxxxxxxx
   21    1.0554 |xxxxxxxxxxxxxxxxxxxxx
   11    1.0877 |xxxxxxxxxxx
    7    1.1200 |xxxxxxx
    7    1.1523 |xxxxxxx
   10    1.1846 |xxxxxxxxxx
    4    1.2169 |xxxx
    3   Extreme | (1.2318) (1.2538) (1.2652)
-----           +---------+---------+---------+---------+---------+
  200           0         10        20        30        40        50

                Histogram estimate G3
```

Figure 4.3: An example of histograms of bootstrap replications.

Appendix A

Technical appendix

In this appendix, the theory of maximum likelihood estimation used in the `MLA` program will be discussed in detail. Other authors, such as Raudenbush and Bryk (2002) and Longford (1987) give some technical detail as well, but much is left to the reader. We think, however, that it is useful to explain in much more detail what is actually done in the program, and this appendix serves this purpose. The opinion of De Leeuw and Kreft (2001) about this appendix (in the previous version of this manual) was:

> "[The manual] is a bit wordy, because it tries to spell out all details, especially the technical ones. The ultimate example is the 20-page Technical Appendix A, which gives in painstaking detail the derivations of the formulae for the likelihood function, its first and second derivatives, and their expectations. It illustrates the effect TEX has on the mind of an individual who has just escaped from the dungeons of WYSIWIG."

Although there is considerable truth in this observation, the primary erason for including these detailed derivations in the first version of this manual was quite different. When we started programming, most authors referred to the technical appendix of (Longford, 1987) for the computational details. When we implemented his formulas, however, our program did not converge. After we spent considerable time in deriving all the formulas, it turned out that there was a typo in one of his formulas. To save others who might want to do the same a lot of time, and to prove that our formulas are correct, we included this appendix.

Our "escape from the dungeons of WYSIWIG" has now been more than 10 years ago, but we still think this technical appendix is useful for (some) users. Apart from the argument above, another important reason is that it occurs frequently that different statistical programs give different answers (estimates, standard errors, test statistics, p-values) in cases where they are expected to do the same analysis. Questions about how this can happen are often found on discussion lists and "frequently asked questions" lists of statistical packages. Such questions can only be answered if it can be assessed what each program does, i.e., which (default) choices are made, which formulas are used, which algorithms are implemented. This appendix provides this information.

In this appendix, the (minus-log-)likelihood function and its gradient function are derived, as well as computationally more efficient formulas of them. The asymptotic covariance matrix of the maximum likelihood estimators and computationally efficient formulas for it are derived and the explicit imposition of implicit constraints in the model is discussed.

A.1 The model and the likelihood function

To find maximum likelihood estimates, we start with the model (2.4):

$$y_j = X_j\gamma + Z_j u_j + \varepsilon_j \tag{A.1}$$

$$\varepsilon_j \sim \mathcal{N}(0, \sigma^2 I_{N_j}) \tag{A.2}$$

$$u_j \sim \mathcal{N}(0, \Theta), \tag{A.3}$$

where y_j is a vector with the endogenous variable for the N_j Level-1 units in Level-2 unit j, X_j is an $N_j \times p$ matrix of exogenous variables for the Level-1 units in Level-2 unit j, and Z_j is an $N_j \times q$ matrix of exogenous variables for the Level-1 units in Level-2 unit j. The p-vector γ is a vector of fixed regression coefficients, the q-vector u_j is a vector of random regression coefficients in Level-2 unit j, and the N_j-vector ε_j is a vector of residuals of the Level-1 units in Level-2 unit j. It is assumed that ε_j and u_j are independent of each other and independent of $\varepsilon_{j'}$ and $u_{j'}$, where $j' \neq j$.

From the model equations (A.1)–(A.3), it is found that, conditional on X_j and Z_j, y_j is normally distributed and the expectation and covariance matrix of y_j are

$$\mathrm{E}\, y_j = X_j\gamma \tag{A.4}$$

$$\begin{aligned} V_j &= \mathrm{E}(y_j - X_j\gamma)(y_j - X_j\gamma)' \\ &= \sigma^2 I_{N_j} + Z_j \Theta Z_j'. \end{aligned} \tag{A.5}$$

Consequently, the probability density of y_j is

$$f(y_j) = \frac{1}{(2\pi)^{N_j/2}(\det V_j)^{1/2}} e^{-\frac{1}{2}(y_j - X_j\gamma)' V_j^{-1}(y_j - X_j\gamma)},$$

so that the contribution of Level-2 unit j to the minus-log-likelihood function is

$$\begin{aligned} L_j &= -\log f(y_j) \\ &= \frac{N_j}{2}\log(2\pi) + \frac{1}{2}\log\det V_j + \frac{1}{2}(y_j - X_j\gamma)' V_j^{-1}(y_j - X_j\gamma) \end{aligned}$$

and the minus-log-likelihood function for the whole sample is

$$\begin{aligned} L &= \sum_{j=1}^{J} L_j \\ &= \frac{N}{2}\log(2\pi) + \frac{1}{2}\sum_{j=1}^{J}\log\det V_j + \frac{1}{2}\sum_{j=1}^{J}(y_j - X_j\gamma)' V_j^{-1}(y_j - X_j\gamma), \end{aligned} \tag{A.6}$$

where J is the number of Level-2 units, and N is the total number of Level-1 units, $N = \sum_{j=1}^{J} N_j$, where N_j is the number of Level-1 units in Level-2 unit j. This is the function that has to be minimized with respect to the parameters to obtain maximum likelihood estimators. To minimize this function, the program uses the *Broyden-Fletcher-Goldfarb-Shanno* (BFGS) minimization method (see, e.g., Nocedal & Wright, 1999), which uses the gradient of the function to be minimized.

In section A.3, computationally efficient formulas for the function and the gradient will be derived. In section A.4, the asymptotic covariance matrix of the estimators will be derived. In section A.5, a reparametrization of the model will be discussed, in which the restriction of positive (semi-)definiteness of covariance matrices is explicitly imposed. But first, in the next section, some useful notation, matrices, and formulas will be introduced.

A.2 Some useful formulas

First, we define the matrix

$$G_j = I_q + Z_j' Z_j \Theta / \sigma^2. \tag{A.7}$$

This matrix will be used frequently in the following.

The inverse of V_j. Wansbeek and Meijer (2000, p. 351) state the following formula:

$$(A + BCD')^{-1} = A^{-1} - A^{-1}B(C^{-1} + D'A^{-1}B)^{-1}D'A^{-1},$$

where A and C are square nonsingular matrices and B and D are matrices of appropriate dimensions. This formula can also be written as

$$\begin{aligned}(A + BCD')^{-1} &= A^{-1} - A^{-1}B[(I + D'A^{-1}BC)C^{-1}]^{-1}D'A^{-1} \\ &= A^{-1} - A^{-1}BC(I + D'A^{-1}BC)^{-1}D'A^{-1}.\end{aligned} \tag{A.8}$$

By defining $A = \sigma^2 I_{N_j}$, $B = Z_j$, $C = \Theta$, and $D = Z_j$, it follows that V_j can be written as $A + BCD'$. Consequently, the inverse of V_j can be found from equation (A.8):

$$\begin{aligned}V_j^{-1} &= \sigma^{-2} I_{N_j} - \left(\sigma^{-2} I_{N_j}\right) Z_j \Theta \left[I_q + Z_j' \left(\sigma^{-2} I_{N_j}\right) Z_j \Theta\right]^{-1} Z_j' \left(\sigma^{-2} I_{N_j}\right) \\ &= \sigma^{-2} I_{N_j} - \sigma^{-4} Z_j \Theta \left(I_q + Z_j' Z_j \Theta / \sigma^2\right)^{-1} Z_j' \\ &= \sigma^{-2} I_{N_j} - \sigma^{-4} Z_j \Theta G_j^{-1} Z_j'.\end{aligned} \tag{A.9}$$

The determinant of V_j. Based on Maddala (1977, pp. 446–447), the following formula for the determinant of a partitioned matrix can be derived:

$$\begin{aligned}\det\begin{pmatrix} A & B \\ C & D \end{pmatrix} &= \det\begin{pmatrix} A & B \\ C & D \end{pmatrix}\begin{pmatrix} I & -A^{-1}B \\ 0 & I \end{pmatrix} \\ &= \det\begin{pmatrix} A & 0 \\ C & D - CA^{-1}B \end{pmatrix} \\ &= \det A \ \det(D - CA^{-1}B),\end{aligned}$$

where A and D are square nonsingular matrices, and B and C are matrices of appropriate orders. Similarly,

$$\begin{aligned}\det\begin{pmatrix} A & B \\ C & D \end{pmatrix} &= \det\begin{pmatrix} A & B \\ C & D \end{pmatrix}\begin{pmatrix} I & 0 \\ -D^{-1}C & I \end{pmatrix} \\ &= \det D \ \det(A - BD^{-1}C).\end{aligned}$$

Consequently,

$$\det A \ \det(D - CA^{-1}B) = \det D \ \det(A - BD^{-1}C). \tag{A.10}$$

Now, define $A = I_q$, $B = Z_j'$, $C = -Z_j\Theta$, and $D = \sigma^2 I_{N_j}$. The matrix V_j can now be written as $V_j = D - CA^{-1}B$, and the determinant of A is 1. Consequently, using equation (A.10),

$$\begin{aligned}
\det V_j &= \det A \det(D - CA^{-1}B) \\
&= \det D \det(A - BD^{-1}C) \\
&= \det(\sigma^2 I_{N_j}) \det[I_q - (Z_j')(\sigma^2 I_{N_j})^{-1}(-Z_j\Theta)] \\
&= (\sigma^2)^{N_j} \det(I_q + Z_j'Z_j\Theta/\sigma^2) \\
&= (\sigma^2)^{N_j} \det G_j. \qquad \text{(A.11)}
\end{aligned}$$

The factor $Z_j'V_j^{-1}$. In the following, the factor $Z_j'V_j^{-1}$ will frequently pop up. This factor can be written in a computationally more efficient form:

$$\begin{aligned}
Z_j'V_j^{-1} &= Z_j'(\sigma^{-2}I_{N_j} - \sigma^{-4}Z_j\Theta G_j^{-1}Z_j') \qquad \text{(from (A.9))} \\
&= \sigma^{-2}Z_j' - \sigma^{-4}Z_j'Z_j\Theta G_j^{-1}Z_j' \\
&= \sigma^{-2}Z_j' - \sigma^{-2}(Z_j'Z_j\Theta/\sigma^2)G_j^{-1}Z_j' \\
&= \sigma^{-2}Z_j' - \sigma^{-2}(I_q + Z_j'Z_j\Theta/\sigma^2 - I_q)G_j^{-1}Z_j' \\
&= \sigma^{-2}Z_j' - \sigma^{-2}(G_j - I_q)G_j^{-1}Z_j' \qquad \text{(from (A.7))} \\
&= \sigma^{-2}Z_j' - \sigma^{-2}Z_j' + \sigma^{-2}G_j^{-1}Z_j' \\
&= \sigma^{-2}G_j^{-1}Z_j'. \qquad \text{(A.12)}
\end{aligned}$$

From (A.12) it follows that

$$Z_j'V_j^{-1}Z_j = \sigma^{-2}G_j^{-1}Z_j'Z_j \qquad \text{(A.13)}$$

and

$$\begin{aligned}
Z_j'V_j^{-2}Z_j &= \sigma^{-2}G_j^{-1}Z_j'V_j^{-1}Z_j \\
&= \sigma^{-4}G_j^{-2}Z_j'Z_j. \qquad \text{(A.14)}
\end{aligned}$$

The traces of V_j^{-1} and V_j^{-2}. From equation (A.9), we find

$$\begin{aligned}
\operatorname{tr} V_j^{-1} &= \operatorname{tr}\left(\sigma^{-2}I_{N_j} - \sigma^{-4}Z_j\Theta G_j^{-1}Z_j'\right) \\
&= \sigma^{-2}N_j - \sigma^{-4}\operatorname{tr}(Z_j\Theta G_j^{-1}Z_j') \\
&= \sigma^{-2}N_j - \sigma^{-4}\operatorname{tr}(Z_j'Z_j\Theta G_j^{-1}) \\
&= \sigma^{-2}N_j - \sigma^{-2}\operatorname{tr}[(Z_j'Z_j\Theta/\sigma^2)G_j^{-1}] \\
&= \sigma^{-2}N_j - \sigma^{-2}\operatorname{tr}[(I_q + Z_j'Z_j\Theta/\sigma^2 - I_q)G_j^{-1}] \\
&= \sigma^{-2}N_j - \sigma^{-2}\operatorname{tr}[(G_j - I_q)G_j^{-1}] \\
&= \sigma^{-2}N_j - \sigma^{-2}\operatorname{tr}(I_q - G_j^{-1}) \\
&= \sigma^{-2}N_j - \sigma^{-2}q + \sigma^{-2}\operatorname{tr} G_j^{-1}. \qquad \text{(A.15)}
\end{aligned}$$

Similarly, using (A.9), (A.15), and (A.12),

$$
\begin{aligned}
\operatorname{tr} V_j^{-2} &= \operatorname{tr}\left[\left(\sigma^{-2} I_{N_j} - \sigma^{-4} Z_j \Theta G_j^{-1} Z_j'\right) V_j^{-1}\right] \\
&= \sigma^{-2} \operatorname{tr} V_j^{-1} - \sigma^{-4} \operatorname{tr}(Z_j \Theta G_j^{-1} Z_j' V_j^{-1}) \\
&= \sigma^{-4}(N_j - q) + \sigma^{-4} \operatorname{tr} G_j^{-1} - \sigma^{-4} \operatorname{tr}\left[Z_j \Theta G_j^{-1}(\sigma^{-2} G_j^{-1} Z_j')\right] \\
&= \sigma^{-4}(N_j - q) + \sigma^{-4} \operatorname{tr} G_j^{-1} - \sigma^{-4} \operatorname{tr}\left\{\left[(Z_j' Z_j \Theta/\sigma^2) G_j^{-1}\right] G_j^{-1}\right\} \\
&= \sigma^{-4}(N_j - q) + \sigma^{-4} \operatorname{tr} G_j^{-1} - \sigma^{-4} \operatorname{tr}\left[(I_q - G_j^{-1}) G_j^{-1}\right] \\
&= \sigma^{-4}(N_j - q) + \sigma^{-4} \operatorname{tr} G_j^{-1} - \sigma^{-4} \operatorname{tr} G_j^{-1} + \sigma^{-4} \operatorname{tr} G_j^{-2} \\
&= \sigma^{-4}(N_j - q) + \sigma^{-4} \operatorname{tr} G_j^{-2}. \qquad \text{(A.16)}
\end{aligned}
$$

Differential formulas. As was stated in the previous section, the maximum likelihood estimates are obtained by minimizing the minus-log-likelihood function by the BFGS method, which uses the gradient of the function. To find the gradient, the differential notation of Magnus and Neudecker (1985, 1988) will be used. The key property of differentials is their relation with derivatives through the following equivalence: Let f be a vector or scalar function of a vector or scalar variable x, then

$$\frac{\partial f}{\partial x'} = A(x) \Leftrightarrow \mathrm{d}\, f = A(x)\ \mathrm{d}\, x.$$

The differential of a matrix is defined through the vector that stacks its columns: $\operatorname{vec} \mathrm{d}\, F = \mathrm{d} \operatorname{vec} F$. Note that the differential of a scalar, vector, or matrix is a scalar, vector, or matrix of the same size.

Some useful formulas are (Magnus & Neudecker, 1985, 1988):

$$
\begin{aligned}
\mathrm{d}(c) &= 0 \\
\mathrm{d}(cg) &= c\ \mathrm{d}\, g \\
\mathrm{d}(g + h) &= \mathrm{d}\, g + \mathrm{d}\, h \\
\mathrm{d}(gh) &= (\mathrm{d}\, g)h + g\ \mathrm{d}\, h \\
\mathrm{d}(\log f) &= \frac{1}{f}\, \mathrm{d}\, f \\
\mathrm{d}(\det F) &= \det F\ \operatorname{tr}(F^{-1}\ \mathrm{d}\, F) \\
\mathrm{d}(\operatorname{tr} F) &= \operatorname{tr}\ \mathrm{d}\, F \\
\mathrm{d}(F^{-1}) &= -F^{-1}(\mathrm{d}\, F)F^{-1},
\end{aligned}
$$

where c is a scalar, vector, or matrix constant, g and h may be scalars, vectors, or matrices (provided the expression is a valid expression), f is a scalar, and F is a matrix.

There is also a *chain rule*: If f is a function of x and g is a function of f, then (cf. Magnus & Neudecker, 1988, p. 91)

$$\frac{\partial g}{\partial x'} = \frac{\partial g}{\partial f'} \frac{\partial f}{\partial x'}.$$

This means the following for the differentials: If $\mathrm{d}\, g = A\ \mathrm{d}\, f$ and $\mathrm{d}\, f = B\ \mathrm{d}\, x$, then $\mathrm{d}\, g = AB\ \mathrm{d}\, x$, which illustrates that, informally speaking, the formulas for the differentials can be filled in sequentially. A similar formula holds for differentials of matrices. In the following, it will be clear how the chain rule can be applied.

The formulas above can be used to derive some important differentials:

$$\begin{aligned}
\mathrm{d}\,V_j &= \mathrm{d}(\sigma^2 I_{N_j} + Z_j \Theta Z_j') \\
&= (\mathrm{d}\,\sigma^2) I_{N_j} + Z_j (\mathrm{d}\,\Theta) Z_j' \qquad (A.17)\\
\mathrm{d}\,\det V_j &= \det V_j \;\operatorname{tr}(V_j^{-1}\;\mathrm{d}\,V_j) \qquad (A.18)\\
\mathrm{d}\;V_j^{-1} &= -V_j^{-1}(\mathrm{d}\,V_j)V_j^{-1} \qquad (A.19)\\
\mathrm{d}\;\log\det V_j &= \frac{1}{\det V_j}\;\mathrm{d}\det V_j \\
&= \frac{1}{\det V_j}\;\det V_j\;\operatorname{tr}(V_j^{-1}\;\mathrm{d}\,V_j) \\
&= \operatorname{tr}(V_j^{-1}\;\mathrm{d}\,V_j). \qquad (A.20)
\end{aligned}$$

Combining equation (A.19) with (A.17), we find that

$$\begin{aligned}
\mathrm{d}\;V_j^{-1} &= -V_j^{-1}[(\mathrm{d}\,\sigma^2)I_{N_j} + Z_j(\mathrm{d}\,\Theta)Z_j']V_j^{-1} \\
&= -V_j^{-2}\,\mathrm{d}\,\sigma^2 - V_j^{-1}Z_j(\mathrm{d}\,\Theta)Z_j'V_j^{-1}. \qquad (A.21)
\end{aligned}$$

Consider a term of the form

$$\mathrm{d}\,T = \operatorname{tr} A\,\mathrm{d}\,\Theta,$$

where Θ is a symmetric $q \times q$ matrix. This term can be written as

$$\begin{aligned}
\mathrm{d}\,T &= \sum_{k=1}^{q}\sum_{l=1}^{q} A_{kl}\,\mathrm{d}\,\Theta_{lk} \\
&= \sum_{k=1}^{q}\sum_{l=1}^{k-1}(A_{kl} + A_{lk})\,\mathrm{d}\,\Theta_{kl} + \sum_{k=1}^{q} A_{kk}\,\mathrm{d}\,\Theta_{kk},
\end{aligned}$$

so

$$\begin{aligned}
\frac{\partial T}{\partial \Theta_{kl}} &= A_{kl} + A_{lk} \qquad (A.22)\\
&= 2A_{kl}, \qquad \text{if } A \text{ is symmetric}, \qquad (A.23)
\end{aligned}$$

and

$$\frac{\partial T}{\partial \Theta_{kk}} = A_{kk}, \qquad (A.24)$$

where $k \neq l$.

Similarly, consider a term of the form

$$\mathrm{d}\,S = [A(\mathrm{d}\,\Theta)B]_{kl},$$

where Θ is a symmetric $q \times q$ matrix, and A and B are matrices. This term can be written as

$$\begin{aligned}
\mathrm{d}\,S &= \sum_{u=1}^{q}\sum_{v=1}^{q} A_{ku}(\mathrm{d}\,\Theta_{uv})B_{vl} \\
&= \sum_{u=1}^{q}\sum_{v=1}^{u-1}(A_{ku}B_{vl} + A_{kv}B_{ul})\,\mathrm{d}\,\Theta_{uv} + \sum_{u=1}^{q} A_{ku}B_{ul}\,\mathrm{d}\,\Theta_{uu},
\end{aligned}$$

so

$$\frac{\partial S}{\partial \Theta_{uv}} = A_{ku}B_{vl} + A_{kv}B_{ul} \tag{A.25}$$

and

$$\frac{\partial S}{\partial \Theta_{uu}} = A_{ku}B_{ul}, \tag{A.26}$$

where $u \neq v$.

A.3 Computational formulas for the function and gradient

The formula (A.6) of the minus-log-likelihood function is computationally inefficient, because a matrix of size N_j has to be inverted, and its determinant calculated. Therefore, in this section a computationally efficient formula will be derived, based on Longford (1987), and using formulas from the previous section. Along the same lines, computationally efficient formulas for the derivatives of this function with respect to the parameters will also be derived.

Combining (A.6), (A.9), and (A.11), we find the following formula for L:

$$\begin{aligned}
L &= \frac{N}{2}\log 2\pi + \frac{1}{2}\sum_{j=1}^{J}\log[(\sigma^2)^{N_j}\det G_j] \\
&\quad + \frac{1}{2}\sum_{j=1}^{J}(y_j - X_j\gamma)'[\sigma^{-2}I_{N_j} - \sigma^{-4}Z_j\Theta G_j^{-1}Z_j'](y_j - X_j\gamma) \\
&= \frac{N}{2}\log 2\pi + \frac{N}{2}\log(\sigma^2) + \frac{1}{2}\sum_{j=1}^{J}\log\det G_j \\
&\quad + \frac{1}{2}\sigma^{-2}\sum_{j=1}^{J}(y_j - X_j\gamma)'(y_j - X_j\gamma) \\
&\quad - \frac{1}{2}\sigma^{-4}\sum_{j=1}^{J}(y_j - X_j\gamma)'Z_j\Theta G_j^{-1}Z_j'(y_j - X_j\gamma) \\
&= \frac{N}{2}\log 2\pi + \frac{N}{2}\log(\sigma^2) + \frac{1}{2}\sum_{j=1}^{J}\log\det G_j \\
&\quad + \frac{1}{2}\sigma^{-2}\left[\left(\sum_{j=1}^{J}y_j'y_j\right) - 2\gamma'\left(\sum_{j=1}^{J}X_j'y_j\right) + \gamma'\left(\sum_{j=1}^{J}X_j'X_j\right)\gamma\right] \\
&\quad - \frac{1}{2}\sigma^{-4}\sum_{j=1}^{J}(Z_j'y_j - Z_j'X_j\gamma)'\Theta G_j^{-1}(Z_j'y_j - Z_j'X_j\gamma).
\end{aligned} \tag{A.27}$$

Formula (A.27) is a computationally efficient formula, and this is the formula that is implemented in the program.

To find the gradient of L, we start with the differential of L:

$$\begin{aligned}\mathrm{d}\,L = \frac{1}{2}\sum_{j=1}^{J} &\mathrm{d}\,\log\,\det V_j \\ &+ \frac{1}{2}\sum_{j=1}^{J} 2(y_j - X_j\gamma)' V_j^{-1}(-X_j\,\mathrm{d}\,\gamma) \\ &+ \frac{1}{2}\sum_{j=1}^{J}(y_j - X_j\gamma)'\,\mathrm{d}(V_j^{-1})(y_j - X_j\gamma). \end{aligned} \tag{A.28}$$

Combining (A.28) with (A.20), (A.21) and (A.17), we find

$$\begin{aligned}\mathrm{d}\,L &= \frac{1}{2}\sum_{j=1}^{J}\mathrm{tr}(V_j^{-1}\,\mathrm{d}\,V_j) - \sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-1}X_j\,\mathrm{d}\,\gamma \\ &\quad - \frac{1}{2}\sum_{j=1}^{J}(y_j - X_j\gamma)'[V_j^{-2}\,\mathrm{d}\,\sigma^2 + V_j^{-1}Z_j(\mathrm{d}\,\Theta)Z_j'V_j^{-1}](y_j - X_j\gamma) \\ &= \frac{1}{2}\sum_{j=1}^{J}\mathrm{tr}\{V_j^{-1}[(\mathrm{d}\,\sigma^2)I_{N_j} + Z_j(\mathrm{d}\,\Theta)Z_j']\} \\ &\quad - \sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-1}X_j\,\mathrm{d}\,\gamma \\ &\quad - \frac{1}{2}\sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-2}(\mathrm{d}\,\sigma^2)(y_j - X_j\gamma) \\ &\quad - \frac{1}{2}\sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-1}Z_j(\mathrm{d}\,\Theta)Z_j'V_j^{-1}(y_j - X_j\gamma) \\ &= \left(\frac{1}{2}\sum_{j=1}^{J}\mathrm{tr}\,V_j^{-1}\right)\mathrm{d}\,\sigma^2 + \frac{1}{2}\sum_{j=1}^{J}\mathrm{tr}[V_j^{-1}Z_j(\mathrm{d}\,\Theta)Z_j'] \\ &\quad - \sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-1}X_j\,\mathrm{d}\,\gamma \\ &\quad - \frac{1}{2}\sum_{j=1}^{J}\left[(y_j - X_j\gamma)'V_j^{-2}(y_j - X_j\gamma)\right]\mathrm{d}\,\sigma^2 \\ &\quad - \frac{1}{2}\sum_{j=1}^{J}\mathrm{tr}\left\{[(y_j - X_j\gamma)'V_j^{-1}Z_j](\mathrm{d}\,\Theta)[Z_j'V_j^{-1}(y_j - X_j\gamma)]\right\}\end{aligned}$$

$$\begin{aligned}
&= \left(\frac{1}{2} \sum_{j=1}^{J} \operatorname{tr} V_j^{-1} \right) \mathrm{d}\,\sigma^2 + \frac{1}{2} \sum_{j=1}^{J} \operatorname{tr}(Z_j' V_j^{-1} Z_j \,\mathrm{d}\,\Theta) \\
&\quad - \left[\sum_{j=1}^{J} (y_j - X_j\gamma)' V_j^{-1} X_j \right] \mathrm{d}\,\gamma \\
&\quad - \left[\frac{1}{2} \sum_{j=1}^{J} (y_j - X_j\gamma)' V_j^{-2} (y_j - X_j\gamma) \right] \mathrm{d}\,\sigma^2 \\
&\quad - \frac{1}{2} \sum_{j=1}^{J} \operatorname{tr}\Big(\big\{ \big[Z_j' V_j^{-1} (y_j - X_j\gamma) \big] \big[Z_j' V_j^{-1} (y_j - X_j\gamma) \big]' \big\} \,\mathrm{d}\,\Theta \Big).
\end{aligned}$$

So

$$\frac{\partial L}{\partial \gamma'} = - \sum_{j=1}^{J} (y_j - X_j\gamma)' V_j^{-1} X_j \tag{A.29}$$

and

$$\frac{\partial L}{\partial \sigma^2} = \frac{1}{2} \sum_{j=1}^{J} \operatorname{tr} V_j^{-1} - \frac{1}{2} \sum_{j=1}^{J} (y_j - X_j\gamma)' V_j^{-2} (y_j - X_j\gamma), \tag{A.30}$$

and, using (A.23) and (A.24),

$$\begin{aligned}
\frac{\partial L}{\partial \Theta_{kl}} &= \sum_{j=1}^{J} (Z_j' V_j^{-1} Z_j)_{kl} \\
&\qquad - \sum_{j=1}^{J} \big\{ \big[Z_j' V_j^{-1} (y_j - X_j\gamma) \big] \big[Z_j' V_j^{-1} (y_j - X_j\gamma) \big]' \big\}_{kl} \\
&= \sum_{j=1}^{J} (Z_j' V_j^{-1} Z_j)_{kl} \\
&\qquad - \sum_{j=1}^{J} \big[Z_j' V_j^{-1} (y_j - X_j\gamma) \big]_k \big[Z_j' V_j^{-1} (y_j - X_j\gamma) \big]_l
\end{aligned} \tag{A.31}$$

and

$$\frac{\partial L}{\partial \Theta_{kk}} = \frac{1}{2} \sum_{j=1}^{J} (Z_j' V_j^{-1} Z_j)_{kk} - \frac{1}{2} \sum_{j=1}^{J} \big[Z_j' V_j^{-1} (y_j - X_j\gamma) \big]_k^2. \tag{A.32}$$

Now, using (A.9), (A.12), (A.13), and (A.15), we find computationally more efficient formulas for the derivatives:

$$
\begin{aligned}
\frac{\partial L}{\partial \gamma} &= -\sum_{j=1}^{J} X_j' V_j^{-1}(y_j - X_j\gamma) \\
&= -\sum_{j=1}^{J} X_j'(\sigma^{-2} I_{N_j} - \sigma^{-4} Z_j \Theta G_j^{-1} Z_j')(y_j - X_j\gamma) \\
&= -\sigma^{-2}\left[\left(\sum_{j=1}^{J} X_j' y_j\right) - \left(\sum_{j=1}^{J} X_j' X_j\right)\gamma\right] \\
&\quad + \sigma^{-4}\sum_{j=1}^{J} X_j' Z_j \Theta G_j^{-1}(Z_j' y_j - Z_j' X_j\gamma), \\
\frac{\partial L}{\partial \sigma^2} &= \frac{1}{2}\sum_{j=1}^{J} \operatorname{tr} V_j^{-1} - \frac{1}{2}\sum_{j=1}^{J}(y_j - X_j\gamma)' V_j^{-2}(y_j - X_j\gamma) \\
&= \frac{1}{2}\sigma^{-2}(N - Jq) + \frac{1}{2}\sigma^{-2}\sum_{j=1}^{J} \operatorname{tr} G_j^{-1} \\
&\quad - \frac{1}{2}\sum_{j=1}^{J}(y_j - X_j\gamma)'(\sigma^{-2} I_{N_j} - \sigma^{-4} Z_j \Theta G_j^{-1} Z_j') V_j^{-1}(y_j - X_j\gamma) \\
&= \frac{1}{2}\sigma^{-2}(N - Jq) + \frac{1}{2}\sigma^{-2}\sum_{j=1}^{J} \operatorname{tr} G_j^{-1} \\
&\quad - \frac{1}{2}\sigma^{-2}\sum_{j=1}^{J}(y_j - X_j\gamma)' V_j^{-1}(y_j - X_j\gamma) \\
&\quad + \frac{1}{2}\sigma^{-4}\sum_{j=1}^{J}(y_j - X_j\gamma)' Z_j \Theta G_j^{-1} Z_j' V_j^{-1}(y_j - X_j\gamma) \\
&= \frac{1}{2}\sigma^{-2}(N - Jq) + \frac{1}{2}\sigma^{-2}\sum_{j=1}^{J} \operatorname{tr} G_j^{-1} \\
&\quad - \frac{1}{2}\sigma^{-2}\sum_{j=1}^{J}(y_j - X_j\gamma)'(\sigma^{-2} I_{N_j} - \sigma^{-4} Z_j \Theta G_j^{-1} Z_j')(y_j - X_j\gamma) \\
&\quad + \frac{1}{2}\sigma^{-4}\sum_{j=1}^{J}(y_j - X_j\gamma)' Z_j \Theta G_j^{-1}(\sigma^{-2} G_j^{-1} Z_j')(y_j - X_j\gamma)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{2}\sigma^{-2}(N - Jq) + \frac{1}{2}\sigma^{-2}\sum_{j=1}^{J} \operatorname{tr} G_j^{-1} \\
&\quad - \frac{1}{2}\sigma^{-4}\left[\left(\sum_{j=1}^{J} y_j'y_j\right) - 2\gamma'\left(\sum_{j=1}^{J} X_j'y_j\right) + \gamma'\left(\sum_{j=1}^{J} X_j'X_j\right)\gamma\right] \\
&\quad + \frac{1}{2}\sigma^{-6}\sum_{j=1}^{J}(Z_j'y_j - Z_j'X_j\gamma)'\Theta G_j^{-1}(Z_j'y_j - Z_j'X_j\gamma) \\
&\quad + \frac{1}{2}\sigma^{-6}\sum_{j=1}^{J}(Z_j'y_j - Z_j'X_j\gamma)'\Theta G_j^{-2}(Z_j'y_j - Z_j'X_j\gamma) \\
&= \frac{1}{2}\sigma^{-2}(N - Jq) + \frac{1}{2}\sigma^{-2}\sum_{j=1}^{J} \operatorname{tr} G_j^{-1} \\
&\quad - \frac{1}{2}\sigma^{-4}\left[\left(\sum_{j=1}^{J} y_j'y_j\right) - 2\gamma'\left(\sum_{j=1}^{J} X_j'y_j\right) + \gamma'\left(\sum_{j=1}^{J} X_j'X_j\right)\gamma\right] \\
&\quad + \frac{1}{2}\sigma^{-6}\sum_{j=1}^{J}(Z_j'y_j - Z_j'X_j\gamma)'\Theta(I_q + G_j^{-1})G_j^{-1}(Z_j'y_j - Z_j'X_j\gamma),
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial L}{\partial \Theta_{kl}} &= \sum_{j=1}^{J}(Z_j'V_j^{-1}Z_j)_{kl} \\
&\quad - \sum_{j=1}^{J}[Z_j'V_j^{-1}(y_j - X_j\gamma)]_k[Z_j'V_j^{-1}(y_j - X_j\gamma)]_l \\
&= \sum_{j=1}^{J}(\sigma^{-2}G_j^{-1}Z_j'Z_j)_{kl} \\
&\quad - \sum_{j=1}^{J}[\sigma^{-2}G_j^{-1}(Z_j'y_j - Z_j'X_j\gamma)]_k[\sigma^{-2}G_j^{-1}(Z_j'y_j - Z_j'X_j\gamma)]_l,
\end{aligned}
$$

and

$$
\frac{\partial L}{\partial \Theta_{kk}} = \frac{1}{2}\sum_{j=1}^{J}(\sigma^{-2}G_j^{-1}Z_j'Z_j)_{kk} - \frac{1}{2}\sum_{j=1}^{J}[\sigma^{-2}G_j^{-1}(Z_j'y_j - Z_j'X_j\gamma)]_k^2.
$$

These formulas are implemented in the program. Note (cf. Bryk & Raudenbush, 1992, p. 239) that the function and the derivatives depend on the data only through the terms $\sum_{j=1}^{J} y_j'y_j$, $\sum_{j=1}^{J} X_j'y_j$, $\sum_{j=1}^{J} X_j'X_j$, $Z_j'y_j$, $Z_j'X_j$, and $Z_j'Z_j$ (through G_j), the first of which is a scalar, the second a p-vector, and the third a symmetric $p \times p$ matrix. The last three are a q-vector, a $q \times p$ matrix, and a symmetric $q \times q$ matrix, for each Level-2 unit. The symmetric matrices may be stored linearly, thereby saving additional memory.

A.4 The asymptotic covariance matrix of the estimators

The asymptotic distribution of the maximum likelihood estimators, under appropriate general conditions is given by (see Magnus, 1978)

$$\sqrt{N}\left(\hat{\theta}_{\mathrm{ML}} - \theta\right) \xrightarrow{d} \mathcal{N}\left[0, \lim_{N\to\infty}\left(\frac{\mathcal{I}(\theta)}{N}\right)^{-1}\right], \tag{A.33}$$

where N is the sample size,

$$\mathcal{I}(\theta) = \mathrm{E}\left(\frac{\partial^2 L}{\partial\theta\,\partial\theta'}\right), \tag{A.34}$$

and L is the minus-log-likelihood function. Therefore, the asymptotic covariance matrix of the estimators is derived from the matrix of second derivatives of L (the Hessian matrix).

From (A.29) we have

$$\frac{\partial L}{\partial \gamma} = -\sum_{j=1}^{J} X_j' V_j^{-1}(y_j - X_j\gamma).$$

Thus,

$$\begin{aligned}
\mathrm{d}\left(\frac{\partial L}{\partial\gamma}\right) &= -\sum_{j=1}^{J} X_j'(\mathrm{d}\,V_j^{-1})(y_j - X_j\gamma) + \sum_{j=1}^{J} X_j' V_j^{-1} X_j \,\mathrm{d}\,\gamma \\
&= \sum_{j=1}^{J} X_j'[V_j^{-2}\,\mathrm{d}\,\sigma^2 + V_j^{-1} Z_j(\mathrm{d}\,\Theta) Z_j' V_j^{-1}](y_j - X_j\gamma) \\
&\quad + \sum_{j=1}^{J} X_j' V_j^{-1} X_j \,\mathrm{d}\,\gamma \qquad \text{(using (A.21))} \\
&= \left[\sum_{j=1}^{J} X_j' V_j^{-2}(y_j - X_j\gamma)\right] \mathrm{d}\,\sigma^2 \\
&\quad + \sum_{j=1}^{J} (X_j' V_j^{-1} Z_j)(\mathrm{d}\,\Theta)[Z_j' V_j^{-1}(y_j - X_j\gamma)] \\
&\quad + \left(\sum_{j=1}^{J} X_j' V_j^{-1} X_j\right) \mathrm{d}\,\gamma.
\end{aligned} \tag{A.35}$$

Therefore,

$$\frac{\partial^2 L}{\partial\gamma\,\partial\gamma'} = \sum_{j=1}^{J} X_j' V_j^{-1} X_j, \tag{A.36}$$

$$\frac{\partial^2 L}{\partial\gamma\,\partial\sigma^2} = \sum_{j=1}^{J} X_j' V_j^{-2}(y_j - X_j\gamma), \tag{A.37}$$

and, using (A.25) and (A.26),

$$\begin{aligned}\frac{\partial^2 L}{\partial\gamma_k\,\partial\Theta_{uv}} &= \sum_{j=1}^{J}\Big\{(X_j'V_j^{-1}Z_j)_{ku}[Z_j'V_j^{-1}(y_j - X_j\gamma)]_v \\ &\qquad + (X_j'V_j^{-1}Z_j)_{kv}[Z_j'V_j^{-1}(y_j - X_j\gamma)]_u\Big\}, \qquad \text{(A.38)}\\ \frac{\partial^2 L}{\partial\gamma_k\,\partial\Theta_{uu}} &= \sum_{j=1}^{J}(X_j'V_j^{-1}Z_j)_{ku}[Z_j'V_j^{-1}(y_j - X_j\gamma)]_u. \qquad \text{(A.39)}\end{aligned}$$

From (A.30), we have

$$\frac{\partial L}{\partial\sigma^2} = \frac{1}{2}\sum_{j=1}^{J}\operatorname{tr}(V_j^{-1}) - \frac{1}{2}\sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-2}(y_j - X_j\gamma).$$

Thus,

$$\begin{aligned}\mathrm{d}\left(\frac{\partial L}{\partial\sigma^2}\right) &= \frac{1}{2}\sum_{j=1}^{J}\operatorname{tr}(\mathrm{d}\,V_j^{-1}) - \frac{1}{2}\sum_{j=1}^{J}(y_j - X_j\gamma)'(\mathrm{d}\,V_j^{-2})(y_j - X_j\gamma) \\ &\qquad + \sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-2}X_j\;\mathrm{d}\,\gamma \\ &= -\frac{1}{2}\sum_{j=1}^{J}\operatorname{tr}[V_j^{-2}\,\mathrm{d}\,\sigma^2 + V_j^{-1}Z_j(\mathrm{d}\,\Theta)Z_j'V_j^{-1}] \\ &\qquad - \frac{1}{2}\sum_{j=1}^{J}(y_j - X_j\gamma)'(\mathrm{d}\,V_j^{-2})(y_j - X_j\gamma) \\ &\qquad + \sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-2}X_j\;\mathrm{d}\,\gamma \quad \text{(using (A.21))} \\ &= \left(-\frac{1}{2}\sum_{j=1}^{J}\operatorname{tr}V_j^{-2}\right)\mathrm{d}\,\sigma^2 - \frac{1}{2}\sum_{j=1}^{J}\operatorname{tr}[(V_j^{-1}Z_j)(\mathrm{d}\,\Theta)(Z_j'V_j^{-1})] \\ &\qquad - \frac{1}{2}\sum_{j=1}^{J}(y_j - X_j\gamma)'[(\mathrm{d}\,V_j^{-1})V_j^{-1} + V_j^{-1}\;\mathrm{d}\,V_j^{-1}](y_j - X_j\gamma) \\ &\qquad + \sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-2}X_j\;\mathrm{d}\,\gamma \\ &= \left(-\frac{1}{2}\sum_{j=1}^{J}\operatorname{tr}V_j^{-2}\right)\mathrm{d}\,\sigma^2 - \frac{1}{2}\sum_{j=1}^{J}\operatorname{tr}[(Z_j'V_j^{-2}Z_j)\,\mathrm{d}\,\Theta] \\ &\qquad - \sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-1}(\mathrm{d}\,V_j^{-1})(y_j - X_j\gamma) \\ &\qquad + \sum_{j=1}^{J}(y_j - X_j\gamma)'V_j^{-2}X_j\;\mathrm{d}\,\gamma\end{aligned}$$

$$
\begin{aligned}
&= \left(-\frac{1}{2} \sum_{j=1}^{J} \operatorname{tr} V_j^{-2} \right) \mathrm{d}\,\sigma^2 - \frac{1}{2} \sum_{j=1}^{J} \operatorname{tr}[(Z_j' V_j^{-2} Z_j)\,\mathrm{d}\,\Theta] \\
&\quad + \sum_{j=1}^{J} (y_j - X_j\gamma)' V_j^{-1} [V_j^{-2}\,\mathrm{d}\,\sigma^2 + V_j^{-1} Z_j (\mathrm{d}\,\Theta) Z_j' V_j^{-1}] (y_j - X_j\gamma) \\
&\quad + \sum_{j=1}^{J} (y_j - X_j\gamma)' V_j^{-2} X_j \;\mathrm{d}\,\gamma \\
&= \left(-\frac{1}{2} \sum_{j=1}^{J} \operatorname{tr} V_j^{-2} \right) \mathrm{d}\,\sigma^2 - \frac{1}{2} \sum_{j=1}^{J} \operatorname{tr}[(Z_j' V_j^{-2} Z_j)\,\mathrm{d}\,\Theta] \\
&\quad + \left[\sum_{j=1}^{J} (y_j - X_j\gamma)' V_j^{-3} (y_j - X_j\gamma) \right] \mathrm{d}\,\sigma^2 \\
&\quad + \sum_{j=1}^{J} [(y_j - X_j\gamma)' V_j^{-2} Z_j](\mathrm{d}\,\Theta)[Z_j' V_j^{-1} (y_j - X_j\gamma)] \\
&\quad + \sum_{j=1}^{J} (y_j - X_j\gamma)' V_j^{-2} X_j \;\mathrm{d}\,\gamma. \\
&= \left(-\frac{1}{2} \sum_{j=1}^{J} \operatorname{tr} V_j^{-2} \right) \mathrm{d}\,\sigma^2 - \frac{1}{2} \sum_{j=1}^{J} \operatorname{tr}[(Z_j' V_j^{-2} Z_j)\,\mathrm{d}\,\Theta] \\
&\quad + \left[\sum_{j=1}^{J} (y_j - X_j\gamma)' V_j^{-3} (y_j - X_j\gamma) \right] \mathrm{d}\,\sigma^2 \\
&\quad + \sum_{j=1}^{J} \operatorname{tr} \left\{ \left[Z_j' V_j^{-1} (y_j - X_j\gamma)(y_j - X_j\gamma)' V_j^{-2} Z_j \right] \mathrm{d}\,\Theta \right\} \\
&\quad + \sum_{j=1}^{J} (y_j - X_j\gamma)' V_j^{-2} X_j \;\mathrm{d}\,\gamma. \qquad \text{(A.40)}
\end{aligned}
$$

Therefore,

$$
\frac{\partial^2 L}{\partial\sigma^2\,\partial\sigma^2} = -\frac{1}{2} \sum_{j=1}^{J} \operatorname{tr} V_j^{-2} + \sum_{j=1}^{J} (y_j - X_j\gamma)' V_j^{-3} (y_j - X_j\gamma), \qquad \text{(A.41)}
$$

and, using (A.22), (A.23) and (A.24),

$$
\begin{aligned}
\frac{\partial^2 L}{\partial\sigma^2\,\partial\Theta_{kl}} &= -\sum_{j=1}^{J} (Z_j' V_j^{-2} Z_j)_{kl} \\
&\quad + \sum_{j=1}^{J} \left\{ \left[Z_j' V_j^{-1} (y_j - X_j\gamma)(y_j - X_j\gamma)' V_j^{-2} Z_j \right]_{kl} \right. \\
&\qquad \left. + \left[Z_j' V_j^{-1} (y_j - X_j\gamma)(y_j - X_j\gamma)' V_j^{-2} Z_j \right]_{lk} \right\}, \qquad \text{(A.42)}
\end{aligned}
$$

$$\frac{\partial^2 L}{\partial\sigma^2\,\partial\Theta_{kk}} = -\frac{1}{2}\sum_{j=1}^{J}(Z_j'V_j^{-2}Z_j)_{kk} + \sum_{j=1}^{J}\left[Z_j'V_j^{-1}(y_j - X_j\gamma)(y_j - X_j\gamma)'V_j^{-2}Z_j\right]_{kk} \tag{A.43}$$

From (A.31) we have

$$\frac{\partial L}{\partial\Theta_{kl}} = \sum_{j=1}^{J}(Z_j'V_j^{-1}Z_j)_{kl} - \sum_{j=1}^{J}\left[Z_j'V_j^{-1}(y_j - X_j\gamma)\right]_k\left[Z_j'V_j^{-1}(y_j - X_j\gamma)\right]_l.$$

Thus,

$$\begin{aligned}
\mathrm{d}\left(\frac{\partial L}{\partial\Theta_{kl}}\right) &= \sum_{j=1}^{J}[Z_j'(\mathrm{d}\,V_j^{-1})Z_j]_{kl} \\
&\quad - \sum_{j=1}^{J}\left[Z_j'(\mathrm{d}\,V_j^{-1})(y_j - X_j\gamma)\right]_k\left[Z_j'V_j^{-1}(y_j - X_j\gamma)\right]_l \\
&\quad - \sum_{j=1}^{J}\left[Z_j'V_j^{-1}(y_j - X_j\gamma)\right]_k\left[Z_j'(\mathrm{d}\,V_j^{-1})(y_j - X_j\gamma)\right]_l \\
&\quad + \sum_{j=1}^{J}\left(Z_j'V_j^{-1}X_j\ \mathrm{d}\,\gamma\right)_k\left[Z_j'V_j^{-1}(y_j - X_j\gamma)\right]_l \\
&\quad + \sum_{j=1}^{J}\left[Z_j'V_j^{-1}(y_j - X_j\gamma)\right]_k\left(Z_j'V_j^{-1}X_j\ \mathrm{d}\,\gamma\right)_l \\
&= -\sum_{j=1}^{J}\left\{Z_j'[V_j^{-2}\,\mathrm{d}\,\sigma^2 + V_j^{-1}Z_j(\mathrm{d}\,\Theta)Z_j'V_j^{-1}]Z_j\right\}_{kl} \\
&\quad + \sum_{j=1}^{J}\left\{Z_j'[V_j^{-2}\,\mathrm{d}\,\sigma^2 + V_j^{-1}Z_j(\mathrm{d}\,\Theta)Z_j'V_j^{-1}](y_j - X_j\gamma)\right\}_k \\
&\qquad \times\left[Z_j'V_j^{-1}(y_j - X_j\gamma)\right]_l \\
&\quad + \sum_{j=1}^{J}\left[Z_j'V_j^{-1}(y_j - X_j\gamma)\right]_k \\
&\qquad \times\left\{Z_j'[V_j^{-2}\,\mathrm{d}\,\sigma^2 + V_j^{-1}Z_j(\mathrm{d}\,\Theta)Z_j'V_j^{-1}](y_j - X_j\gamma)\right\}_l \\
&\quad + \sum_{j=1}^{J}\left(Z_j'V_j^{-1}X_j\ \mathrm{d}\,\gamma\right)_k\left[Z_j'V_j^{-1}(y_j - X_j\gamma)\right]_l \\
&\quad + \sum_{j=1}^{J}\left[Z_j'V_j^{-1}(y_j - X_j\gamma)\right]_k\left(Z_j'V_j^{-1}X_j\ \mathrm{d}\,\gamma\right)_l \quad \text{(using (A.21))}
\end{aligned}$$

$$
\begin{aligned}
&= \left[- \sum_{j=1}^{J} (Z_j' V_j^{-2} Z_j)_{kl} \right] \mathrm{d}\,\sigma^2 - \sum_{j=1}^{J} \left[(Z_j' V_j^{-1} Z_j)(\mathrm{d}\,\Theta)(Z_j' V_j^{-1} Z_j) \right]_{kl} \\
&\quad + \left\{ \sum_{j=1}^{J} \left[Z_j' V_j^{-2} (y_j - X_j \gamma) \right]_k \left[Z_j' V_j^{-1} (y_j - X_j \gamma) \right]_l \right\} \mathrm{d}\,\sigma^2 \\
&\quad + \sum_{j=1}^{J} \Big\{ (Z_j' V_j^{-1} Z_j)(\mathrm{d}\,\Theta) \\
&\qquad\quad \times \left[Z_j' V_j^{-1} (y_j - X_j \gamma)(y_j - X_j \gamma)' V_j^{-1} Z_j \right] \Big\}_{kl} \\
&\quad + \left\{ \sum_{j=1}^{J} \left[Z_j' V_j^{-1} (y_j - X_j \gamma) \right]_k \left[Z_j' V_j^{-2} (y_j - X_j \gamma) \right]_l \right\} \mathrm{d}\,\sigma^2 \\
&\quad + \sum_{j=1}^{J} \Big\{ \left[Z_j' V_j^{-1} (y_j - X_j \gamma)(y_j - X_j \gamma)' V_j^{-1} Z_j \right] \\
&\qquad\quad \times (\mathrm{d}\,\Theta)(Z_j' V_j^{-1} Z_j) \Big\}_{kl} \\
&\quad + \sum_{j=1}^{J} \left(Z_j' V_j^{-1} X_j \;\mathrm{d}\,\gamma \right)_k \left[Z_j' V_j^{-1} (y_j - X_j \gamma) \right]_l \\
&\quad + \sum_{j=1}^{J} \left[Z_j' V_j^{-1} (y_j - X_j \gamma) \right]_k \left(Z_j' V_j^{-1} X_j \;\mathrm{d}\,\gamma \right)_l
\end{aligned}
\tag{A.44}
$$

Combining (A.44) with (A.25) and (A.26), the partial derivatives are found:

$$
\begin{aligned}
\frac{\partial^2 L}{\partial \Theta_{kl}\, \partial \Theta_{uv}} &= - \sum_{j=1}^{J} \left[(Z_j' V_j^{-1} Z_j)_{ku} (Z_j' V_j^{-1} Z_j)_{vl} + (Z_j' V_j^{-1} Z_j)_{kv} (Z_j' V_j^{-1} Z_j)_{ul} \right] \\
&\quad + \sum_{j=1}^{J} \Big\{ (Z_j' V_j^{-1} Z_j)_{ku} \left[Z_j' V_j^{-1} (y_j - X_j \gamma)(y_j - X_j \gamma)' V_j^{-1} Z_j \right]_{vl} \\
&\qquad\quad + (Z_j' V_j^{-1} Z_j)_{kv} \left[Z_j' V_j^{-1} (y_j - X_j \gamma)(y_j - X_j \gamma)' V_j^{-1} Z_j \right]_{ul} \Big\} \\
&\quad + \sum_{j=1}^{J} \Big\{ \left[Z_j' V_j^{-1} (y_j - X_j \gamma)(y_j - X_j \gamma)' V_j^{-1} Z_j \right]_{ku} (Z_j' V_j^{-1} Z_j)_{vl} \\
&\qquad\quad + \left[Z_j' V_j^{-1} (y_j - X_j \gamma)(y_j - X_j \gamma)' V_j^{-1} Z_j \right]_{kv} (Z_j' V_j^{-1} Z_j)_{ul} \Big\} \\
&= - \sum_{j=1}^{J} \left[(Z_j' V_j^{-1} Z_j)_{ku} (Z_j' V_j^{-1} Z_j)_{vl} + (Z_j' V_j^{-1} Z_j)_{kv} (Z_j' V_j^{-1} Z_j)_{ul} \right] \\
&\quad + \sum_{j=1}^{J} \Big\{ (Z_j' V_j^{-1} Z_j)_{ku} \left[Z_j' V_j^{-1} (y_j - X_j \gamma)(y_j - X_j \gamma)' V_j^{-1} Z_j \right]_{vl} \\
&\qquad\quad + (Z_j' V_j^{-1} Z_j)_{kv} \left[Z_j' V_j^{-1} (y_j - X_j \gamma)(y_j - X_j \gamma)' V_j^{-1} Z_j \right]_{ul} \\
&\qquad\quad + (Z_j' V_j^{-1} Z_j)_{ul} \left[Z_j' V_j^{-1} (y_j - X_j \gamma)(y_j - X_j \gamma)' V_j^{-1} Z_j \right]_{kv} \\
&\qquad\quad + (Z_j' V_j^{-1} Z_j)_{vl} \left[Z_j' V_j^{-1} (y_j - X_j \gamma)(y_j - X_j \gamma)' V_j^{-1} Z_j \right]_{ku} \Big\}
\end{aligned}
\tag{A.45}
$$

and

$$\begin{aligned}
\frac{\partial^2 L}{\partial \Theta_{kl}\,\partial \Theta_{uu}} &= -\sum_{j=1}^{J}\left[(Z_j' V_j^{-1} Z_j)_{ku}(Z_j' V_j^{-1} Z_j)_{ul}\right] \\
&\quad + \sum_{j=1}^{J}\left\{(Z_j' V_j^{-1} Z_j)_{ku}\left[Z_j' V_j^{-1}(y_j - X_j\gamma)(y_j - X_j\gamma)' V_j^{-1} Z_j\right]_{ul}\right\} \\
&\quad + \sum_{j=1}^{J}\left\{\left[Z_j' V_j^{-1}(y_j - X_j\gamma)(y_j - X_j\gamma)' V_j^{-1} Z_j\right]_{ku}(Z_j' V_j^{-1} Z_j)_{ul}\right\} \\
&= -\sum_{j=1}^{J}\left[(Z_j' V_j^{-1} Z_j)_{ku}(Z_j' V_j^{-1} Z_j)_{ul}\right] \\
&\quad + \sum_{j=1}^{J}\left\{(Z_j' V_j^{-1} Z_j)_{ku}\left[Z_j' V_j^{-1}(y_j - X_j\gamma)(y_j - X_j\gamma)' V_j^{-1} Z_j\right]_{ul}\right. \\
&\qquad \left. + \left[Z_j' V_j^{-1}(y_j - X_j\gamma)(y_j - X_j\gamma)' V_j^{-1} Z_j\right]_{ku}(Z_j' V_j^{-1} Z_j)_{ul}\right\}. \qquad \text{(A.46)}
\end{aligned}$$

Analogously, from (A.32) we have

$$\frac{\partial L}{\partial \Theta_{kk}} = \frac{1}{2}\sum_{j=1}^{J}(Z_j' V_j^{-1} Z_j)_{kk} - \frac{1}{2}\sum_{j=1}^{J}\left[Z_j' V_j^{-1}(y_j - X_j\gamma)\right]_k^2,$$

and

$$\begin{aligned}
\mathrm{d}\left(\frac{\partial L}{\partial \Theta_{kk}}\right) &= \left[-\frac{1}{2}\sum_{j=1}^{J}(Z_j' V_j^{-2} Z_j)_{kk}\right]\mathrm{d}\,\sigma^2 \\
&\quad - \frac{1}{2}\sum_{j=1}^{J}\left[(Z_j' V_j^{-1} Z_j)(\mathrm{d}\,\Theta)(Z_j' V_j^{-1} Z_j)\right]_{kk} \\
&\quad + \left\{\frac{1}{2}\sum_{j=1}^{J}\left[Z_j' V_j^{-2}(y_j - X_j\gamma)\right]_k\left[Z_j' V_j^{-1}(y_j - X_j\gamma)\right]_k\right\}\mathrm{d}\,\sigma^2 \\
&\quad + \frac{1}{2}\sum_{j=1}^{J}\left\{(Z_j' V_j^{-1} Z_j)(\mathrm{d}\,\Theta)\right. \\
&\qquad \left. \times\left[Z_j' V_j^{-1}(y_j - X_j\gamma)(y_j - X_j\gamma)' V_j^{-1} Z_j\right]\right\}_{kk} \\
&\quad + \left\{\frac{1}{2}\sum_{j=1}^{J}\left[Z_j' V_j^{-1}(y_j - X_j\gamma)\right]_k\left[Z_j' V_j^{-2}(y_j - X_j\gamma)\right]_k\right\}\mathrm{d}\,\sigma^2 \\
&\quad + \frac{1}{2}\sum_{j=1}^{J}\left\{\left[Z_j' V_j^{-1}(y_j - X_j\gamma)(y_j - X_j\gamma)' V_j^{-1} Z_j\right]\right. \\
&\qquad \left. \times(\mathrm{d}\,\Theta)(Z_j' V_j^{-1} Z_j)\right\}_{kk} \\
&\quad + \frac{1}{2}\sum_{j=1}^{J}(Z_j' V_j^{-1} X_j\ \mathrm{d}\,\gamma)_k\left[Z_j' V_j^{-1}(y_j - X_j\gamma)\right]_k \\
&\quad + \frac{1}{2}\sum_{j=1}^{J}\left[Z_j' V_j^{-1}(y_j - X_j\gamma)\right]_k(Z_j' V_j^{-1} X_j\ \mathrm{d}\,\gamma)_k. \qquad \text{(A.47)}
\end{aligned}$$

Combining (A.47) with (A.26) we have

$$\begin{aligned}\frac{\partial^2 L}{\partial \Theta_{kk}\,\partial \Theta_{uu}} &= -\frac{1}{2}\sum_{j=1}^{J}(Z_j' V_j^{-1} Z_j)_{ku}^2 \\ &\quad + \frac{1}{2}\sum_{j=1}^{J}(Z_j' V_j^{-1} Z_j)_{ku} \\ &\qquad \times \left[Z_j' V_j^{-1}(y_j - X_j\gamma)(y_j - X_j\gamma)' V_j^{-1} Z_j\right]_{uk} \\ &\quad + \frac{1}{2}\sum_{j=1}^{J}\left[Z_j' V_j^{-1}(y_j - X_j\gamma)(y_j - X_j\gamma)' V_j^{-1} Z_j\right]_{ku} \\ &\qquad \times (Z_j' V_j^{-1} Z_j)_{uk} \\ &= -\frac{1}{2}\sum_{j=1}^{J}(Z_j' V_j^{-1} Z_j)_{ku}^2 \\ &\quad + \sum_{j=1}^{J}(Z_j' V_j^{-1} Z_j)_{ku} \\ &\qquad \times \left[Z_j' V_j^{-1}(y_j - X_j\gamma)(y_j - X_j\gamma)' V_j^{-1} Z_j\right]_{ku}, \end{aligned} \tag{A.48}$$

because the matrices between brackets and parentheses in (A.48) are symmetric.

Now, from (A.33) and (A.34), we have to take expectations of the second derivatives. Therefore, from (A.4) and (A.5), the following expectations will be used:

$$\begin{aligned} \mathrm{E}(y_j - X_j\gamma) &= 0 \\ \mathrm{E}(y_j - X_j\gamma)(y_j - X_j\gamma)' &= V_j. \end{aligned}$$

From (A.36), (A.37), (A.38), and (A.39), we have

$$\mathrm{E}\left(\frac{\partial^2 L}{\partial \gamma\,\partial \gamma'}\right) = \sum_{j=1}^{J} X_j' V_j^{-1} X_j, \tag{A.49}$$

$$\mathrm{E}\left(\frac{\partial^2 L}{\partial \gamma\,\partial \sigma^2}\right) = 0, \tag{A.50}$$

$$\mathrm{E}\left(\frac{\partial^2 L}{\partial \gamma\,\partial \Theta_{uv}}\right) = 0, \tag{A.51}$$

and

$$\mathrm{E}\left(\frac{\partial^2 L}{\partial \gamma\,\partial \Theta_{uu}}\right) = 0. \tag{A.52}$$

From (A.41), (A.42), and (A.43), we have

$$\begin{aligned}\mathrm{E}\left(\frac{\partial^2 L}{\partial \sigma^2\,\partial \sigma^2}\right) &= -\frac{1}{2}\sum_{j=1}^{J}\operatorname{tr} V_j^{-2} + \sum_{j=1}^{J}\mathrm{E}\left[(y_j - X_j\gamma)' V_j^{-3}(y_j - X_j\gamma)\right] \\ &= -\frac{1}{2}\sum_{j=1}^{J}\operatorname{tr} V_j^{-2} + \sum_{j=1}^{J}\mathrm{E}\ \operatorname{tr}\left[V_j^{-3}(y_j - X_j\gamma)(y_j - X_j\gamma)'\right]\end{aligned}$$

$$= \frac{1}{2} \sum_{j=1}^{J} \operatorname{tr} V_j^{-2}; \tag{A.53}$$

$$\begin{aligned}\frac{\partial^2 L}{\partial \sigma^2 \, \partial \Theta_{kl}} &= - \sum_{j=1}^{J} (Z_j' V_j^{-2} Z_j)_{kl} \\ &\quad + \sum_{j=1}^{J} \mathrm{E} \Big\{ \big[Z_j' V_j^{-1} (y_j - X_j\gamma)(y_j - X_j\gamma)' V_j^{-2} Z_j \big]_{kl} \\ &\qquad + \big[Z_j' V_j^{-1} (y_j - X_j\gamma)(y_j - X_j\gamma)' V_j^{-2} Z_j \big]_{lk} \Big\}, \\ &= \sum_{j=1}^{J} (Z_j' V_j^{-2} Z_j)_{kl};\end{aligned} \tag{A.54}$$

and

$$\begin{aligned}\mathrm{E}\left(\frac{\partial^2 L}{\partial \sigma^2 \, \partial \Theta_{kk}} \right) &= -\frac{1}{2} \sum_{j=1}^{J} (Z_j' V_j^{-2} Z_j)_{kk} \\ &\quad + \sum_{j=1}^{J} \mathrm{E} \big[Z_j' V_j^{-1} (y_j - X_j\gamma)(y_j - X_j\gamma)' V_j^{-2} Z_j \big]_{kk} \\ &= \frac{1}{2} \sum_{j=1}^{J} (Z_j' V_j^{-2} Z_j)_{kk}.\end{aligned} \tag{A.55}$$

From (A.45), (A.46), and (A.48), we have

$$\begin{aligned}\mathrm{E}\left(\frac{\partial^2 L}{\partial \Theta_{kl} \, \partial \Theta_{uv}} \right) &= - \sum_{j=1}^{J} \big[(Z_j' V_j^{-1} Z_j)_{ku} (Z_j' V_j^{-1} Z_j)_{vl} \\ &\qquad + (Z_j' V_j^{-1} Z_j)_{kv} (Z_j' V_j^{-1} Z_j)_{ul} \big] \\ &\quad + \sum_{j=1}^{J} \big[(Z_j' V_j^{-1} Z_j)_{ku} (Z_j' V_j^{-1} Z_j)_{vl} \\ &\qquad + (Z_j' V_j^{-1} Z_j)_{kv} (Z_j' V_j^{-1} Z_j)_{ul} \\ &\qquad + (Z_j' V_j^{-1} Z_j)_{ul} (Z_j' V_j^{-1} Z_j)_{kv} \\ &\qquad + (Z_j' V_j^{-1} Z_j)_{vl} (Z_j' V_j^{-1} Z_j)_{ku} \big] \\ &= \sum_{j=1}^{J} \big[(Z_j' V_j^{-1} Z_j)_{ku} (Z_j' V_j^{-1} Z_j)_{vl} \\ &\qquad + (Z_j' V_j^{-1} Z_j)_{kv} (Z_j' V_j^{-1} Z_j)_{ul} \big],\end{aligned} \tag{A.56}$$

and, analogously,

$$\mathrm{E}\left(\frac{\partial^2 L}{\partial \Theta_{kl} \, \partial \Theta_{uu}} \right) = \sum_{j=1}^{J} (Z_j' V_j^{-1} Z_j)_{ku} (Z_j' V_j^{-1} Z_j)_{ul}, \tag{A.57}$$

and

$$\mathrm{E}\left(\frac{\partial^2 L}{\partial \Theta_{kk} \, \partial \Theta_{uu}} \right) = \frac{1}{2} \sum_{j=1}^{J} (Z_j' V_j^{-1} Z_j)_{ku}^2. \tag{A.58}$$

As with the function and the gradient, computationally more efficient formulas will be derived for the covariance matrix of the estimators.

Combining (A.49), (A.50), (A.51), and (A.52) with (A.9), it is found that

$$\begin{aligned}
\mathrm{E}\left(\frac{\partial^2 L}{\partial\gamma\,\partial\gamma'}\right) &= \sum_{j=1}^{J} X_j'\left(\sigma^{-2} I_{N_j} - \sigma^{-4} Z_j \Theta G_j^{-1} Z_j'\right) X_j \\
&= \sigma^{-2}\left(\sum_{j=1}^{J} X_j' X_j\right) - \sigma^{-4}\sum_{j=1}^{J} X_j' Z_j \Theta G_j^{-1} Z_j' X_j \\
\mathrm{E}\left(\frac{\partial^2 L}{\partial\gamma\,\partial\sigma^2}\right) &= 0, \\
\mathrm{E}\left(\frac{\partial^2 L}{\partial\gamma\,\partial\Theta_{uv}}\right) &= 0,
\end{aligned}$$

and

$$\mathrm{E}\left(\frac{\partial^2 L}{\partial\gamma\,\partial\Theta_{uu}}\right) = 0.$$

Combining (A.53), (A.54), and (A.55) with (A.16) and (A.14), it is found that

$$\begin{aligned}
\mathrm{E}\left(\frac{\partial^2 L}{\partial\sigma^2\,\partial\sigma^2}\right) &= \frac{1}{2}\sum_{j=1}^{J}\sigma^{-4}(N_j - q) + \frac{1}{2}\sum_{j=1}^{J}\sigma^{-4}\,\mathrm{tr}\,G_j^{-2} \\
&= \frac{1}{2}\sigma^{-4}(N - Jq) + \frac{1}{2}\sigma^{-4}\sum_{j=1}^{J}\mathrm{tr}\,G_j^{-2}; \\
\mathrm{E}\left(\frac{\partial^2 L}{\partial\sigma^2\,\partial\Theta_{kl}}\right) &= \sigma^{-4}\sum_{j=1}^{J}(G_j^{-2} Z_j' Z_j)_{kl};
\end{aligned}$$

and

$$\mathrm{E}\left(\frac{\partial^2 L}{\partial\sigma^2\,\partial\Theta_{kk}}\right) = \frac{1}{2}\sigma^{-4}\sum_{j=1}^{J}(G_j^{-2} Z_j' Z_j)_{kk}.$$

Combining (A.56), (A.57), and (A.58) with (A.13), it is found that

$$\begin{aligned}
\mathrm{E}\left(\frac{\partial^2 L}{\partial\Theta_{kl}\,\partial\Theta_{uv}}\right) &= \sigma^{-4}\sum_{j=1}^{J}\Big[(G_j^{-1} Z_j' Z_j)_{ku}(G_j^{-1} Z_j' Z_j)_{vl} \\
&\qquad\qquad + (G_j^{-1} Z_j' Z_j)_{kv}(G_j^{-1} Z_j' Z_j)_{ul}\Big]; \\
\mathrm{E}\left(\frac{\partial^2 L}{\partial\Theta_{kl}\,\partial\Theta_{uu}}\right) &= \sigma^{-4}\sum_{j=1}^{J}(G_j^{-1} Z_j' Z_j)_{ku}(G_j^{-1} Z_j' Z_j)_{ul}; \\
\mathrm{E}\left(\frac{\partial^2 L}{\partial\Theta_{kk}\,\partial\Theta_{uu}}\right) &= \frac{1}{2}\sigma^{-4}\sum_{j=1}^{J}(G_j^{-1} Z_j' Z_j)_{ku}^2.
\end{aligned}$$

These formulas are implemented in the program. Note that these expressions depend on the data only through the terms $\sum_{j=1}^{J} X_j' X_j$, $Z_j' X_j$, and $Z_j' Z_j$, which are also used for the function and gradient (cf. section A.3), so that no additional memory is required for data storage.

Let H be the matrix defined by these expressions. Then

$$\frac{H}{N} \xrightarrow{p} \lim_{N\to\infty} \left(\frac{\mathcal{I}(\theta)}{N} \right),$$

where $\mathcal{I}(\theta)$ is given by equation (A.34), and θ is the parameter vector that has to be estimated. So $(H/N)^{-1}$ is a consistent estimator of the asymptotic covariance matrix of $\sqrt{N}(\hat{\theta} - \theta)$, or H^{-1} is the estimator of the covariance matrix of $\hat{\theta}$.

A.5 Reparametrization

In the formulas of the previous sections, all parameters were treated as completely free parameters. But σ^2 should obviously be nonnegative, because it is a variance. Similarly, Θ should be a positive (semi-)definite matrix, because it is a covariance matrix.

As discussed in chapter 3, `MLA` offers two options for reparameterization to impose these restrictions: "root" and "logarithmic". Corresponding formulas will be derived in a next version of this manual. Here, an old description of a previous reparameterization method, based on the Cholesky decomposition and used in earlier versions of `MLA` is still given instead.

The parameters can be written in the following way:

$$\sigma^2 = (\sigma)^2 \tag{A.59}$$

$$\Theta = CC', \tag{A.60}$$

where C is a lower triangular matrix (i.e., with zero elements above the diagonal). Equation (A.59) states that σ should be the parameter used by the program, not σ^2. Equation (A.60) expresses Θ in its Cholesky decomposition, and the elements of C should be the parameters used by the program. This reparametrization may have some drawbacks (cf. Gill, Murray, & Wright, 1981, pp. 268–269), but we think that it may generally be useful for multilevel analysis. See also Longford (1987), who uses a similar reparametrization of a restricted model. Note that the reparametrization (A.60) cannot be easily used if some elements of Θ are restricted.

In order to minimize the reparametrized function, the gradient vector should be reparametrized accordingly. This is done by using the chain rule of partial derivatives: If the original parameter vector is denoted by θ, and the reparametrized parameter vector by ϕ, then

$$\frac{\partial L}{\partial \phi'} = \frac{\partial L}{\partial \theta'} \frac{\partial \theta}{\partial \phi'}. \tag{A.61}$$

Therefore, the formulas from section A.3 have to be postmultiplied by

$$\frac{\partial \theta}{\partial \phi'}.$$

The relevant formula for σ is

$$\frac{\partial \sigma^2}{\partial \sigma} = 2\sigma.$$

To form the relevant expression for C, consider the (k,l) and (k,k) elements of $\varTheta$, where $k > l$:

$$\begin{aligned}\varTheta_{kl} &= \sum_{u=1}^{q} C_{ku} C_{lu} \\ &= \sum_{u=1}^{l} C_{ku} C_{lu}; \\ \varTheta_{kk} &= \sum_{u=1}^{q} C_{ku}^2 \\ &= \sum_{u=1}^{k} C_{ku}^2.\end{aligned}$$

So,

$$\begin{aligned}
\frac{\partial \varTheta_{kl}}{\partial C_{ku}} &= C_{lu}, && \text{if } u \le l; \\
\frac{\partial \varTheta_{kl}}{\partial C_{ku}} &= 0, && \text{if } u > l; \\
\frac{\partial \varTheta_{kl}}{\partial C_{lu}} &= C_{ku}, && \text{if } u \le l; \\
\frac{\partial \varTheta_{kl}}{\partial C_{lu}} &= 0, && \text{if } u > l; \\
\frac{\partial \varTheta_{kl}}{\partial C_{uv}} &= 0, && \text{if } u \ne k \text{ and } u \ne l; \\
\frac{\partial \varTheta_{kk}}{\partial C_{ku}} &= 2C_{ku}, && \text{if } u \le k; \\
\frac{\partial \varTheta_{kk}}{\partial C_{ku}} &= 0, && \text{if } u > k; \\
\frac{\partial \varTheta_{kk}}{\partial C_{uv}} &= 0, && \text{if } u \ne k.
\end{aligned}$$

Consequently, if $u \ge v$,

$$\begin{aligned}\frac{\partial L}{\partial C_{uv}} &= \sum_{k=1}^{q} \sum_{l=1}^{k-1} \frac{\partial L}{\partial \varTheta_{kl}} \frac{\partial \varTheta_{kl}}{\partial C_{uv}} + \sum_{k=1}^{q} \frac{\partial L}{\partial \varTheta_{kk}} \frac{\partial \varTheta_{kk}}{\partial C_{uv}} \\ &= \sum_{l=v}^{u-1} \frac{\partial L}{\partial \varTheta_{ul}} C_{lv} + \sum_{k=u+1}^{q} \frac{\partial L}{\partial \varTheta_{ku}} C_{kv} + 2 \frac{\partial L}{\partial \varTheta_{uu}} C_{uv}.\end{aligned}$$

These formulas are implemented in the program.

It is possible to transform the second derivatives in a similar way to obtain an estimator of the covariance matrix of the estimators. But, in general, the user will be more interested in the original parameters, and therefore, the estimates of the transformed parameters are retransformed to estimates of the original parameters, and the covariance matrix of section A.4 is used. This procedure is correct, because the transformation is a one-to-one mapping from the feasible region of the original parameters to the domain of the transformed parameters (except for some trivial equivalent solutions, such as σ and $-\sigma$, which lead to the same retransformed solution). Only when the estimates are near the boundary of the feasible region, the asymptotic covariance matrix may not be correct, but the usual statistical theory only applies to interior points, so boundary solutions are a problem in any parametrization.

References

Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society A*, *149*, 1–43. (with discussion)

Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*. New York: Wiley.

Blalock, H. M. (1984). Contextual-effects models: Theoretical and methodological issues. *Annual Review of Sociology*, *10*, 353–372.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Busing, F. M. T. A. (1993). *Distribution characteristics of variance estimates in two-level models; A Monte Carlo study* (Tech. Rep. No. PRM 93-04). Leiden, The Netherlands: Leiden University, Department of Psychometrics and Research Methodology.

Busing, F. M. T. A., Meijer, E., & Van der Leeden, R. (1994). *MLA: Software for multilevel analysis of data with two levels. User's guide for version 1.0b* (Tech. Rep. No. PRM 94-01). Leiden: Leiden University, Department of Psychology.

Busing, F. M. T. A., Meijer, E., & Van der Leeden, R. (1999). Delete-*m* jackknife for unequal *m*. *Statistics and Computing*, *9*, 3–8.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge, UK: Cambridge University Press.

De Leeuw, J. (2005a). Centering in multilevel analysis. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 1, pp. 247–249). New York: Wiley.

De Leeuw, J. (2005b). Linear multilevel analysis. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 2, pp. 1054–1061). New York: Wiley.

De Leeuw, J., & Kreft, I. G. G. (1986). Random coefficient models for multilevel analysis. *Journal of Educational Statistics*, *11*, 57–85.

De Leeuw, J., & Kreft, I. G. G. (1995). Questioning multilevel models. *Journal of Educational and Behavioral Statistics*, *20*, 171–189.

De Leeuw, J., & Kreft, I. G. G. (2001). Software for multilevel analysis. In A. H. Leyland & H. Goldstein (Eds.), *Multilevel modelling of health statistics* (pp. 187–204). Chichester: Wiley.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, *39*, 1–38.

Durbin, J. (1973). *Distribution theory for tests based on the sample distribution function*. Philadelphia: SIAM.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*, 1–26.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Philadelphia: SIAM.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*, 171–200. (with discussion)

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.

Gill, P. E., Murray, W., & Wright, M. H. (1981). *Practical optimization*. London: Academic Press.

Glasnapp, D. R., & Poggio, J. P. (1985). *Essentials of statistical analysis for the behavioral sciences*. Columbus, OH: Charles Merrill.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, *73*, 43–56.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Arnold.

Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics*, *1*, 223–231.

Gong, G., & Samaniego, F. J. (1981). Pseudo maximum likelihood estimation: Theory and applications. *The Annals of Statistics*, *9*, 861–869.

Hall, P. (1992). *The bootstrap and Edgeworth expansion*. New York: Springer.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, *72*, 320–340.

Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Kreft, I. G. G. (1996). *Are multilevel techniques necessary? an overview, including simulation studies.* Retrieved August 22, 2005, from `http://www.calstatela.edu/faculty/ikreft/quarterly/quarterly.html`

Kreft, I. G. G., & De Leeuw, J. (1991). Model based ranking of schools. *International Journal of Educational Research*, *15*, 45–59.

Kreft, I. G. G., & De Leeuw, J. (1998). *Introducing multilevel modelling*. London: Sage.

Kreft, I. G. G., De Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, *30*, 1–21.

Kreft, I. G. G., & Van der Leeden, R. (1994). *Random coefficient linear regression models* (Tech. Rep. No. PRM-03-94). Leiden, The Netherlands: Leiden University, Department of Psychometrics and Research Methodology.

Langer, W. (2004). *Mehrebenenanalyse: eine Einführung für Forschung und Praxis*. Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

LeBlond, D. J. (2005). *Methodology for predicting batch manufacturing risk.* Unpublished master's thesis, Colorado State University, Fort Collins.

Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, *74*, 817–827.

Longford, N. T. (1990). *VARCL. Software for variance component analysis of data with nested random effects (maximum likelihood).* Princeton, NJ: Educational Testing Service.

Maddala, G. S. (1977). *Econometrics*. Singapore: McGraw-Hill.

Magnus, J. R. (1978). Maximum likelihood estimation of the GLS model with unknown parameters in the disturbance covariance matrix. *Journal of Econometrics*, *7*, 281–312.

Magnus, J. R., & Neudecker, H. (1985). Matrix differential calculus with applications to simple, Hadamard and Kronecker products. *Journal of Mathematical Psychology*, *29*, 474–492.

Magnus, J. R., & Neudecker, H. (1988). *Matrix differential calculus with applications in statistics and econometrics*. Chichester: Wiley.

Markus, M. T. (1994). *Bootstrap confidence regions in nonlinear multivariate analysis*. Leiden: DSWO Press.

Mason, W. M., Wong, G. Y., & Entwisle, B. (1983). Contextual analysis through the multilevel linear model. In S. Leinhardt (Ed.), *Sociological methodology 1983 – 1984* (pp. 72–103). San Francisco: Jossey-Bass.

Matsumoto, M., & Nishimura, T. (1998). Mersenne Twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, *8*, 3–30.

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the theory of statistics* (3rd ed.). Singapore: McGraw-Hill.

Murphy, K. M., & Topel, R. H. (1985). Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics*, *3*, 370–379.

Nishimura, T., & Matsumoto, M. (2002). A C-program for MT19937, with initialization improved 2002/1/26 [Computer software]. Retrieved August 29, 2005, from `http://www.math.sci.hiroshima-u.ac.jp/~m-mat/MT/emt.html`

Nocedal, J., & Wright, S. J. (1999). *Numerical optimization*. New York: Springer.

Parke, W. R. (1986). Pseudo maximum likelihood estimation: The asymptotic distribution. *The Annals of Statistics*, *14*, 355–357.

Putter, H. (1994). *Consistency of resampling methods*. Unpublished doctoral dissertation, Leiden University, Leiden, The Netherlands.

Quenouille, M. H. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society B*, *11*, 18–84.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, *43*, 353–360.

Rasbash, J., Steele, F., Browne, W., & Prosser, B. (2004). *A user's guide to MLwiN. Version 2.0.* London: Multilevel Models Project, Institute of Education, University of London.

Raudenbush, S. W. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics*, *13*, 85–116.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. (2004). *HLM 6: Hierarchical linear and nonlinear modeling*. Chicago: Scientific Software International.

Schucany, W. R., Gray, H. L., & Owen, D. B. (1971). On bias reduction in estimation. *Journal of the American Statistical Association*, *66*, 524–533.

Shao, J., & Tu, D. (1995). *The jackknife and bootstrap*. New York: Springer.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.

Stephens, M. A. (1974). EDF statistics for goodness of fit. *Journal of the American Statistical Association*, *69*, 730–737.

Stevens, J. P. (1990). *Intermediate statistics: A modern approach*. Hillsdale, NJ: Erlbaum.

Strenio, J. F., Weisberg, H. I., & Bryk, A. S. (1983). Empirical Bayes estimation of individual growth-curve parameters and their relationship to covariates. *Biometrics*, *39*, 71–86.

Tu, D., & Zhang, L. (1992). Jackknife approximations for some nonparametric confidence intervals of functional parameters based on normalizing transformations. *Computational Statistics*, *7*, 3–15.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples. *The Annals of Mathematical Statistics*, *29*, 614.

Van der Leeden, R. (1998). Multilevel analysis of repeated measures data. *Quality & Quantity*, *32*, 15–29.

Van der Leeden, R., & Busing, F. M. T. A. (1994). *First iteration versus final IGLS/RIGLS estimates in two-level models: A Monte Carlo study with ML3* (Tech. Rep. No. PRM-02-94).

Leiden, The Netherlands: Leiden University, Department of Psychometrics and Research Methodology.

Van Landeghem, G., Onghena, P., & Van Damme, J. (2001). *The effect of different forms of centering in hierarchical linear models re-examined* (Tech. Rep. No. 2001-04). Leuven, Belgium: Catholic University of Leuven, University Centre for Statistics.

Wansbeek, T., & Meijer, E. (2000). *Measurement error and latent variables in econometrics*. Amsterdam: North-Holland.

White, H. L. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1–25.

Wolter, K. M. (1985). *Introduction to variance estimation*. New York: Springer.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, *14*, 1261–1350. (with discussion)