# Multilevel Analysis

## Techniques and Applications

QUANTITATIVE METHODOLOGY SERIES

Second Edition

Joop J. Hox

# Multilevel Analysis

## Second Edition

# QUANTITATIVE METHODOLOGY SERIES

*George A. Marcoulides, Series Editor*

This series presents methodological techniques to investigators and students. The goal is to provide an understanding and working knowledge of each method with a minimum of mathematical derivations. Each volume focuses on a specific method (e.g., Factor Analysis, Multilevel Analysis, Structural Equation Modeling).

Proposals are invited from interested authors. Each proposal should consist of: a brief description of the volume's focus and intended market; a table of contents with an outline of each chapter; and a curriculum vitae. Materials may be sent to Dr. George A. Marcoulides, University of California, Riverside, george.marcoulides@ucr.edu

***Published titles***

**Marcoulides** • *Modern Methods for Business Research*

**Marcoulides & Moustaki** • *Latent Variable and Latent Structure Models*

**van der Ark, Croon, & Sijtsma (Eds.)** • *New Developments in Categorical Data Analysis for the Social and Behavioral Sciences*

**Duncan, Duncan, & Strycker** • *An Introduction to Latent Variable Growth Curve Modeling: Concepts, Issues, and Applications, Second Edition*

**Heck & Thomas** • *An Introduction to Multilevel Modeling Techniques, Second Edition*

**Cardinet, Johnson, & Pini** • *Applying Generalizability Theory Using EduG*

**Heck, Thomas, & Tabata** • *Multilevel and Longitudinal Modeling with PASW/SPSS*

**Hox** • *Multilevel Analysis: Techniques and Applications, Second Edition*

# Multilevel Analysis

## Techniques and Applications

**Second Edition**

**Joop J. Hox**
*Utrecht University, The Netherlands*

Routledge
Taylor & Francis Group
NEW YORK AND HOVE

Published in 2010
by Routledge
270 Madison Avenue
New York, NY 10016
www.psypress.com
www.researchmethodsarena.com

Published in Great Britain
by Routledge
27 Church Road
Hove, East Sussex BN3 2FA

*Routledge is an imprint of the Taylor & Francis Group, an Informa business*

This edition published in the Taylor & Francis e-Library, 2010.

To purchase your own copy of this or any of Taylor & Francis or Routledge's
collection of thousands of eBooks please go to www.eBookstore.tandf.co.uk.

Library of Congress Cataloging-in-Publication Data
Hox, J. J.
Multilevel analysis : techniques and applications / Joop J. Hox. – 2nd ed.
p. cm. – (Quantitative methodology series)
Includes bibliographical references and index.
ISBN 978–1–84872–845–5 (hardback : alk. paper) – ISBN 978–1–84872–846–2 (pbk. : alk. paper) – ISBN 978–0–203–85227–9 (electronic book) 1. Social sciences – Statistical methods. 2. Analysis of variance. 3. Regression analysis. I. Title.
HA29.H783 2010
001.4'22 – dc22
2009048988

ISBN 0-203-85227-3 Master e-book ISBN

ISBN: 978–1–84872–845–5 (hbk)
ISBN: 978–1–84872–846–2 (pbk)

# Contents

# Preface

*To err is human, to forgive divine;*
*but to include errors into your design is statistical.*

—Leslie Kish

This book is intended as an introduction to multilevel analysis for students and applied researchers. The term 'multilevel' refers to a hierarchical or nested data structure, usually subjects within organizational groups, but the nesting may also consist of repeated measures within subjects, or respondents within clusters, as in cluster sampling. The expression *multilevel model* is used as a generic term for all models for nested data. *Multilevel analysis* is used to examine relations between variables measured at different levels of the multilevel data structure. This book presents two types of multilevel models in detail: the multilevel regression model and the multilevel structural equation model.

Multilevel modeling used to be only for specialists. However, in the past decade, multilevel analysis software has become available that is both powerful and accessible. In addition, several books have been published, including the first edition of this book. There is a continuing surge of interest in multilevel analysis, as evidenced by the appearance of several reviews and monographs, applications in different fields ranging from psychology and sociology, to education and medicine, and a thriving Internet discussion list with more than 1400 subscribers. The view of 'multilevel analysis' applying to individuals nested within groups has changed to a view that multilevel models and analysis software offer a very flexible way to model complex data. Thus, multilevel modeling has contributed to the analysis of traditional individuals within groups data, repeated measures and longitudinal data, sociometric modeling, twin studies, meta-analysis and analysis of cluster randomized trials.

In addition to it being an introduction, this book includes a discussion of many extensions and special applications. As an introduction, it is useable in courses on multilevel modeling in a variety of fields, such as psychology, education, sociology, and business. The various extensions and special applications also make it useful to researchers who work in applied or theoretical research, and to methodologists who have to consult with these researchers. The basic models and examples are discussed in non-technical terms; the emphasis is on understanding the methodological and statistical issues involved in using these models. They assume that readers have a basic knowledge of social science statistics, including analysis of variance and multiple regression analysis. The section about multilevel structural equation models assumes a basic understanding of ordinary structural equation modeling. Some of the extensions and special applications contain discussions that are more technical, either because that is necessary for understanding what the model does, or as a helpful introduction to more advanced treatments in other texts. Thus, in addition to its role as an introduction, the book should be useful as a standard reference for a large variety of applications. The chapters that discuss specialized problems, such as the chapter on cross-classified data, the meta-analysis chapter, and the chapter on advanced issues in estimation and testing, can be skipped entirely if preferred.

## New to This Edition

Compared to the first edition, some chapters have changed much, while other chapters have mostly been updated to reflect recent developments in statistical research and software development. One important development is new and better estimation methods for non-normal data that use numerical integration. These are more accurate than the earlier methods that were based on linearization of the non-linear model. These estimation methods have been added to existing software (such as HLM) or are included in more recent software packages (such as SuperMix, Mplus, and the multilevel logistic regression modules in SAS and STATA). The chapter on multilevel logistic regression, and the new chapter on multilevel ordered regression, now include a full treatment of this estimation method. In multilevel structural equation modeling (MSEM) the developments have been so fast that the chapter on multilevel confirmatory factor analysis is completely rewritten, while the chapter on multilevel path analysis is significantly revised. The introduction of the basic two-level regression model in chapter two now also discusses three-level models with an example. The chapter on longitudinal modeling is expanded with a better discussion of covariance structures across time, varying measurement occasions, and a discussion of analyzing data where no growth curve or trend is expected. Some simpler examples have been added to help the novice, but the more complex examples that combine more than one problem have been retained.

Two new chapters were added, one on multilevel models for ordinal and count data, and one on multilevel survival analysis.

An updated website at http://www.joophox.net features data sets for all the text examples formatted using the latest versions of SPSS, HLM, MlwiN, Lisrel, and Mplus, updated screen shots for each of these programs, and PowerPoint slides for instructors. Most analyses in this book can be carried out by any program, although the majority of the multilevel regression analyses were carried out in HLM and MLwiN and the multilevel SEM analyses use LISREL and M*plus*. System files and setups using these packages will also be made available at the website.

Some of the example data are real, while others have been simulated especially for this book. The data sets are quite varied so as to appeal to those in several disciplines, including education, sociology, psychology, family studies, medicine, and nursing, Appendix A describes the various data sets used in this book in detail. In time, further example data will be added to the website for use in computer labs.

**Acknowledgments**

I thank Peter van der Heijden, Herbert Hoijtink, Bernet Sekasanvu Kato, Edith de Leeuw, George Marcoulides, Mirjam Moerbeek, Ian Plewis, Ken Rowe, Godfried van den Wittenboer, and Bill Yeaton for their comments on the manuscript of the first edition of this book. Their critical comments still shape this book. The second edition has profited from comments by Lawrence DeCarlo, Brian Gray, Ellen Hamaker, Don Hedeker, Cora Maas, Cameron McIntosh, Herb Marsh, Mirjam Moerbeek, Allison O'Mara, Elif Unal, and numerous students, as well as our reviewers Dick Carpenter of the University of Colorado, Donald Hedeker of the University of Illinois at Chicago, and others who wish to remain anonymous. And special thanks to the Editor of the Quantitative Methodology Series, George Marcoulides of the University of California – Riverside.

I thank my colleagues at the Department of Methodology and Statistics of the Faculty of Social Sciences at Utrecht University for providing me with many discussions and a generally stimulating research environment. My research has also benefited from the lively discussions by the denizens of the Internet *Multilevel Modeling* and the *Structural Equations Modeling (SEMNET)* discussion lists.

As always, any errors remaining in the book are entirely my own responsibility. I appreciate hearing about them, and will keep a list of errata on the homepage of this book.

J.J. Hox

Amsterdam

# 1

# Introduction to Multilevel Analysis

Social research regularly involves problems that investigate the relationship between individuals and society. The general concept is that individuals interact with the social contexts to which they belong, that individual persons are influenced by the social groups or contexts to which they belong, and that those groups are in turn influenced by the individuals who make up that group. The individuals and the social groups are conceptualized as a hierarchical system of individuals nested within groups, with individuals and groups defined at separate levels of this hierarchical system. Naturally, such systems can be observed at different hierarchical levels, and variables may be defined at each level. This leads to research into the relationships between variables characterizing individuals and variables characterizing groups, a kind of research that is generally referred to as *multilevel research*.

In multilevel research, the data structure in the population is hierarchical, and the sample data are a sample from this hierarchical population. Thus, in educational research, the population consists of schools and pupils within these schools, and the sampling procedure often proceeds in two stages: First, we take a sample of schools, and next we take a sample of pupils within each school. Of course, in real research one may have a convenience sample at either level, or one may decide not to sample pupils but to study all available pupils in the sample of schools. Nevertheless, one should keep firmly in mind that the central statistical model in multilevel analysis is one of successive sampling from each level of a hierarchical population.

In this example, pupils are *nested* within schools. Other examples are cross-national studies where the individuals are nested within their national units, organizational research with individuals nested within departments within organizations, family research with family members within families, and methodological research into interviewer effects with respondents nested within interviewers. Less obvious applications of multilevel models are longitudinal research and growth curve research, where a series of several distinct observations are viewed as nested within individuals, and meta-analysis where the subjects are nested within different studies. For simplicity, this book describes the multilevel models mostly in terms of individuals nested within groups, but note that the models apply to a much larger class of analysis problems.

## 1.1 AGGREGATION AND DISAGGREGATION

In multilevel research, variables can be defined at any level of the hierarchy. Some of these variables may be measured directly at their ‘own’ natural level; for example, at the school level we may measure school size and denomination, and at the pupil level intelligence and school success. In addition, we may move variables from one level to another by aggregation or disaggregation. Aggregation means that the variables at a lower level are moved to a higher level, for instance by assigning to the schools the school mean of the pupils’ intelligence scores. Disaggregation means moving variables to a lower level, for instance by assigning to all pupils in the schools a variable that indicates the denomination of the school they belong to.

The lowest level (level 1) is usually defined by the individuals. However, this is not always the case. Galtung (1969), for instance, defines roles within individuals as the lowest level, and in longitudinal designs, repeated measures within individuals are the lowest level.

At each level in the hierarchy, we may have several types of variables. The distinctions made in the following are based on the typology offered by Lazarsfeld and Menzel (1961), with some simplifications. In our typology, we distinguish between *global*, *structural* and *contextual* variables.

*Global* variables are variables that refer only to the level at which they are defined, without reference to other units or levels. A pupil’s intelligence or gender would be a global variable at the pupil level. School size would be a global variable at the school level. A global variable is measured at the level at which that variable actually exists.

*Structural* variables are operationalized by referring to the sub-units at a lower level. They are constructed from variables at a lower level, for example, in defining the school variable ‘mean intelligence’ as the mean of the intelligence scores of the pupils in that school. Using the mean of a lower-level variable as an explanatory variable at a higher level is a common procedure in multilevel analysis. Other functions of the lower-level variables are less common, but may also be valuable. For instance, using the standard deviation of a lower-level variable as an explanatory variable at a higher level could be used to test hypotheses about the effect of group heterogeneity on the outcome variable. Klein and Kozlowski (2000) refer to such variables as configural variables, and emphasize the importance of capturing the pattern of individual variation in a group. Their examples also emphasize the use of other functions than the mean of individual scores to reflect group characteristics.

It is clear that constructing a structural variable from the lower-level data involves aggregation. *Contextual* variables, on the other hand, refer to the super-units; all units at the lower level receive the value of a variable for the super-unit to which they belong at the higher level. For instance, we can assign to all pupils in a school the school size, or the mean intelligence, as a pupil-level variable. This is called *disaggregation*; data on higher-level units are disaggregated into data on a larger number of lower-level units. The resulting variable is called a *contextual* variable, because it refers to the higher-level context of the units we are investigating.

In order to analyze multilevel models, it is not important to assign each variable to its proper place in the typology. The benefit of the scheme is conceptual; it makes clear to which level a measurement properly belongs. Historically, multilevel problems have led to analysis approaches that moved all variables by aggregation or disaggregation to one single level of interest followed by an ordinary multiple regression, analysis of variance, or some other 'standard' analysis method. However, analyzing variables from different levels at one single common level is inadequate, and leads to two distinct types of problems.

The first problem is statistical. If data are aggregated, the result is that different data values from many sub-units are combined into fewer values for fewer higher-level units. As a result, much information is lost, and the statistical analysis loses power. On the other hand, if data are disaggregated, the result is that a few data values from a small number of super-units are 'blown up' into many more values for a much larger number of sub-units. Ordinary statistical tests treat all these disaggregated data values as independent information from the much larger sample of sub-units. The proper sample size for these variables is of course the number of higher-level units. Using the larger number of disaggregated cases for the sample size leads to significance tests that reject the null-hypothesis far more often than the nominal alpha level suggests. In other words: investigators come up with many 'significant' results that are totally spurious.

The second problem is conceptual. If the analyst is not very careful in the interpretation of the results, s/he may commit the fallacy of the wrong level, which consists of analyzing the data at one level, and formulating conclusions at another level. Probably the best-known fallacy is the *ecological fallacy*, which is interpreting aggregated data at the individual level. It is also known as the 'Robinson effect' after Robinson (1950). Robinson presents aggregated data describing the relationship between the percentage of blacks and the illiteracy level in nine geographic regions in 1930. The *ecological correlation*, that is, the correlation between the aggregated variables at the region level, is 0.95. In contrast, the individual-level correlation between these global variables is 0.20. Robinson concludes that in practice an ecological correlation is almost certainly not equal to its corresponding individual-level correlation. For a statistical explanation, see Robinson (1950) or Kreft and de Leeuw (1987). Formulating inferences at a higher level based on analyses performed at a lower level is just as misleading. This fallacy is known as the *atomistic fallacy*. A related but different fallacy is known as 'Simpson's paradox' (see Lindley & Novick, 1981). Simpson's paradox refers to the problem that completely erroneous conclusions may be drawn if grouped data, drawn from heterogeneous populations, are collapsed and analyzed as if they came from a single homogeneous population. An extensive typology of such fallacies is given by Alker (1969). When aggregated data are the only available data, King (1997) presents some procedures that make it possible to estimate the corresponding individual relationships without committing an ecological fallacy.

A better way to look at multilevel data is to realize that there is not one 'proper' level at which the data should be analyzed. Rather, all levels present in the data are important in their own way. This becomes clear when we investigate cross-level hypotheses, or *multilevel* problems. A multilevel problem is a problem that concerns the relationships between variables that are measured at a number of different hierarchical levels. For example, a common question is how a number of individual and group variables influence one single individual outcome variable. Typically, some of the higher-level explanatory variables may be the aggregated group means of lower-level individual variables. The goal of the analysis is to determine the direct effect of individual- and group-level explanatory variables, and to determine if the explanatory variables at the group level serve as moderators of individual-level relationships. If group-level variables moderate lower-level relationships, this shows up as a statistical interaction between explanatory variables from different levels. In the past, such data were usually analyzed using conventional multiple regression analysis with one dependent variable at the lowest (individual) level and a collection of explanatory variables from all available levels (see Boyd & Iversen, 1979; van den Eeden & Hüttner, 1982). Since this approach analyzes all available data at one single level, it suffers from all of the conceptual and statistical problems mentioned above.

## 1.2 WHY DO WE NEED SPECIAL MULTILEVEL ANALYSIS TECHNIQUES?

A multilevel problem concerns a population with a hierarchical structure. A sample from such a population can be described as a multistage sample: First, we take a sample of units from the higher level (e.g., schools), and next we sample the sub-units from the available units (e.g., we sample pupils from the schools). In such samples, the individual observations are in general not completely independent. For instance, pupils in the same school tend to be similar to each other, because of selection processes (for instance, some schools may attract pupils from higher social economic status (SES) levels, while others attract lower SES pupils) and because of the common history the pupils share by going to the same school. As a result, the average correlation (expressed in the so-called *intraclass correlation*) between variables measured on pupils from the same school will be higher than the average correlation between variables measured on pupils from different schools. Standard statistical tests lean heavily on the assumption of independence of the observations. If this assumption is violated (and in multilevel data this is almost always the case) the estimates of the standard errors of conventional statistical tests are much too small, and this results in many spuriously 'significant' results. The effect is generally *not* negligible, as small dependencies in combination with large group sizes still result in large biases in the standard errors. The strong biases that may be the effect of violation of the assumption of independent observations made in standard statistical tests have been known for a long time (Walsh, 1947) and are still a very important assumption to check in statistical analyses (Stevens, 2009).

The problem of dependencies between individual observations also occurs in survey research, if the sample is not taken at random but cluster sampling from geographical areas is used instead. For similar reasons as in the school example given above, respondents from the same geographical area will be more similar to each other than are respondents from different geographical areas. This leads again to estimates for standard errors that are too small and produce spurious 'significant' results. In survey research, this effect of cluster sampling is well known (see Kish, 1965, 1987). It is called a 'design effect', and various methods are used to deal with it. A convenient correction procedure is to compute the standard errors by ordinary analysis methods, estimate the intraclass correlation between respondents within clusters, and finally employ a correction formula to the standard errors. A correction described by Kish (1965, p. 259) corrects the sampling variance using $v_{eff} = v(1 + (n_{clus} - 1)\rho)$, where $v_{eff}$ is the effective sampling variance, $v$ is the sampling variance calculated by standard methods assuming simple random sampling, $n_{clus}$ is the cluster size, and $\rho$ is the intraclass correlation. The corrected standard error is then equal to the square root of the effective sampling variance The intraclass correlation can be estimated using the between and within mean square from a one-way analysis of variance with the groups as a factor:

$$\rho = (MS_B - MS_W)/(MS_B + (n_{clus} - 1)MS_W).$$

The formula assumes equal group sizes, which is not always realistic. Chapter 2 presents a multilevel model that estimates the intraclass correlation (ICC) without assuming equal group sizes. A variation of the Kish formula computes the effective sample size in two-stage cluster sampling as $n_{eff} = n/[1 + (n_{clus} - 1)\,\rho]$, where $n$ is the total sample size and $n_{eff}$ is the effective sample size. Using this formula, we can simply calculate the effective sample size for different situations, and use weighting to correct the sample size determined by traditional software.[1] For instance, suppose that we take a sample of 10 classes, each with 20 pupils. This comes to a total sample size of 200, which is

[1] The formulas given here apply to two-stage cluster sampling. Other sampling schemes, such as stratified sampling, require different formulas. See Kish (1965, 1987) for details. The symbol $\rho$ (the Greek letter rho) was introduced by Kish (1965, p. 161) who called it *roh* for 'rate of homogeneity'.

reasonable. Let us further suppose that we are interested in a variable for which the intraclass correlation $\rho$ is .10. This seems a rather low intraclass correlation. However, the effective sample size in this situation is 200/[1 + (20 − 1).1] = 69.0, which is much less than the apparent total sample size of 200! Gulliford, Ukoumunne, and Chin (1999) give an overview of estimates of the intraclass correlation to aid in the design of complex health surveys. Their data include variables on a range of lifestyle risk factors and health outcomes, for respondents clustered at the household, postal code, and health authority district levels. They report between-cluster variation at each of these levels, with intraclass correlations ranging from .0 to .3 at the household level, and being mostly smaller than .05 at the postal code level, and below .01 at the district level. Smeeth and Ng (2002) present ICCs for health-related variables for elderly patients within primary-care clinics. Their ICCs are generally small, the largest being .06 for 'difficult to keep house warm'. Smeeth and Ng (2002) list 17 other studies that report ICCs in the field of health research.

Since the design effect depends on both the intraclass correlation and the cluster size, large intraclass correlations are partly compensated by small group sizes. Conversely, small intraclass correlations at the higher levels are offset by the usually large cluster sizes at these levels. Groves (1989) also discusses the effects of cluster sampling on the standard errors in cluster samples, and concludes that the intraclass correlation is usually small, but in combination with the usual cluster sizes used in surveys they still can lead to substantial design effects.

Some of the correction procedures developed for cluster and other complex samples are quite powerful (see Skinner, Holt, & Smith, 1989). In principle such correction procedures could also be applied in analyzing multilevel data, by adjusting the standard errors of the statistical tests. However, multilevel models are multivariate models, and in general the intraclass correlation and hence the effective $N$ is different for different variables. In addition, in most multilevel problems we do not only have clustering of individuals within groups, but we also have variables measured at all available levels, and we are interested in the relationships between all these variables. Combining variables from different levels in one statistical model is a different and more complicated problem than estimating and correcting for design effects. Multilevel models are designed to analyze variables from different levels simultaneously, using a statistical model that properly includes the various dependencies.

To provide an example of a clearly multilevel problem, consider the 'frog pond' theory that has been utilized in educational and organizational research. The 'frog pond' theory refers to the notion that a specific individual frog may be a medium sized frog in a pond otherwise filled with large frogs, or a medium sized frog in a pond otherwise filled with small frogs. Applied to education, this metaphor points out that the effect of an explanatory variable such as 'intelligence' on school career may depend on the average intelligence of the other pupils in the school. A moderately intelligent pupil in a highly intelligent context may become demotivated and thus become an underachiever, while the same pupil in a considerably less intelligent context may gain confidence and become an overachiever. Thus, the effect of an individual pupil's intelligence depends on the average intelligence of the other pupils in the class. A popular approach in educational research to investigate 'frog pond' effects has been to aggregate variables like the pupils' IQs into group means, and then to disaggregate these group means again to the individual level. As a result, the data file contains both individual-level (global) variables and higher-level (contextual) variables in the form of disaggregated group means. Cronbach (1976; Cronbach & Webb, 1979) has suggested expressing the individual scores as deviations from their respective group means, a procedure that has become known as *centering on the group mean*, or *group mean centering*. Centering on the group mean makes very explicit that the individual scores should be interpreted relative to their group's mean. The example of the 'frog pond' theory and the corresponding practice of centering the predictor variables makes clear that combining and analyzing information from different levels within one statistical model is central to multilevel modeling.

## 1.3 MULTILEVEL THEORIES

Multilevel problems must be explained by multilevel theories, an area that seems underdeveloped compared to the advances made in modeling and computing machinery (see Hüttner & van den Eeden, 1993). Multilevel models in general require that the grouping criterion is clear, and that variables can be assigned unequivocally to their appropriate level. In reality, group boundaries are sometimes fuzzy and somewhat arbitrary, and the assignment of variables is not always obvious and simple. In multilevel problems, decisions about group membership and operationalizations involve a wide range of theoretical assumptions, and an equally wide range of specification problems for the auxiliary theory (Blalock, 1990; Klein & Kozlowski, 2000). If there are effects of the social context on individuals, these effects must be mediated by intervening processes that depend on characteristics of the social context. When the number of variables at the different levels is large, there are an enormous number of possible cross-level interactions. Ideally, a multilevel theory should specify which variables belong to which level, and which direct effects and cross-level interaction effects can be expected. Cross-level interaction effects between the individual and the context level require the specification of processes within individuals that cause those individuals to be differentially influenced by certain aspects of the context. Attempts to identify such processes have been made by, among others, Stinchcombe (1968), Erbring and Young (1979), and Chan (1998). The common core in these theories is that they all postulate one or more psychological processes that mediate between individual variables and group variables. Since a global explanation by 'group telepathy' is generally not acceptable, communication processes and the internal structure of groups become important concepts. These are often measured as a 'structural variable'. In spite of their theoretical relevance, structural variables are infrequently used in multilevel research. Another theoretical area that has been largely neglected by multilevel researchers is the influence of individuals on the group. This is already visible in Durkheim's concept of sociology as a science that focuses primarily on the constraints that a society can put on its members, and disregards the influence of individuals on their society. In multilevel modeling, the focus is on models where the outcome variable is at the lowest level. Models that investigate the influence of individual variables on group outcomes are scarce. For a review of this issue see DiPrete and Forristal (1994); an example is discussed by Alba and Logan (1992). Croon and van Veldhoven (2007) discuss analysis methods for multilevel data where the outcome variable is at the highest level.

## 1.4 MODELS DESCRIBED IN THIS BOOK

This book treats two classes of multilevel models: multilevel regression models, and multilevel models for covariance structures.

*Multilevel regression models* are essentially a multilevel version of the familiar multiple regression model. As Cohen and Cohen (1983), Pedhazur (1997), and others have shown, the multiple regression model is very versatile. Using dummy coding for categorical variables, it can be used to analyze analysis of variance (ANOVA)-type of models as well as the more usual multiple regression models. Since the multilevel regression model is an extension of the classical multiple regression model, it too can be used in a wide variety of research problems.

Chapter 2 of this book contains a basic introduction to the multilevel regression model, also known as the hierarchical linear model, or the random coefficient model. Chapters 3 and 4 discuss estimation procedures, and a number of important methodological and statistical issues. They also discuss some technical issues that are not specific to multilevel regression analysis, such as centering and interpreting interactions.

Chapter 5 introduces the multilevel regression model for longitudinal data. The model is a straightforward extension of the standard multilevel regression model, but there are some specific complications, such as autocorrelated errors, which are discussed.

Chapter 6 treats the generalized linear model for dichotomous data and proportions. When the response (dependent) variable is dichotomous or a proportion, standard regression models should not be used. This chapter discusses the multilevel version of the logistic and the probit regression model.

Chapter 7 extends the generalized linear model introduced in Chapter 6 to analyze data that are ordered categorical and to data that are counts. In the context of counts, it presents models that take an overabundance of zeros into account.

Chapter 8 introduces multilevel modeling of survival or event history data. Survival models are for data where the outcome is the occurrence or nonoccurrence of a certain event, in a certain observation period. If the event has not occurred when the observation period ends, the outcome is said to be censored, since we do not know whether or not the event has taken place after the observation period ended.

Chapter 9 discusses cross-classified models. Some data are multilevel in nature, but do not have a neat hierarchical structure. Examples are longitudinal school research data, where pupils are nested within schools, but may switch to a different school in later measurements, and sociometric choice data. Multilevel models for such cross-classified data can be formulated, and estimated with standard software provided that it can handle restrictions on estimated parameters.

Chapter 10 discusses multilevel regression models for multivariate outcomes. These can also be used to estimate models that resemble confirmative factor analysis, and to assess the reliability of multilevel measurements. A different approach to multilevel confirmative factor analysis is treated in Chapter 13.

Chapter 11 describes a variant of the multilevel regression model that can be used in meta-analysis. It resembles the weighted regression model often recommended for meta-analysis. Using standard multilevel regression procedures, it is a flexible analysis tool, especially when the meta-analysis includes multivariate outcomes.

Chapter 12 deals with the sample size needed for multilevel modeling, and the problem of estimating the power of an analysis given a specific sample size. An obvious complication in multilevel power analysis is that there are different sample sizes at the distinct levels, which should be taken into account.

Chapter 13 treats some advanced methods of estimation and assessing significance. It discusses the profile likelihood method, robust standard errors for establishing confidence intervals, and multilevel bootstrap methods for estimating bias-corrected point-estimates and confidence intervals. This chapter also contains an introduction into Bayesian (MCMC) methods for estimation and inference.

*Multilevel models for covariance structures*, or multilevel structural equation models (SEM), are a powerful tool for the analysis of multilevel data. Recent versions of structural equation modeling software, such as EQS, LISREL, Mplus, all include at least some multilevel features. The general statistical model for multilevel covariance structure analysis is quite complicated. Chapter 14 in this book describes both a simplified statistical model proposed by Muthén (1990, 1994), and more recent developments. It explains how multilevel confirmatory factor models can be estimated with either conventional SEM software or using specialized programs. In addition, it deals with issues of calculating standardized coefficients and goodness-of-fit indices in multilevel structural models. Chapter 15 extends this to path models. Chapter 16 describes structural models for latent curve analysis. This is a SEM approach to analyzing longitudinal data, which is very similar to the multilevel regression models treated in Chapter 5.

This book is intended as an introduction to the world of multilevel analysis. Most of the chapters on multilevel regression analysis should be readable for social scientists who have a good general knowledge of analysis of variance and classical multiple regression analysis. Some of these chapters contain material that is more difficult, but this is generally a discussion of specialized problems, which can be skipped at first reading. An example is the chapter on longitudinal models, which contains a prolonged discussion of techniques to model specific structures for the covariances between adjacent time points. This discussion is not needed to understand the essentials of multilevel analysis of longitudinal data, but it may become important when one is actually analyzing such data. The chapters on multilevel structure equation modeling obviously require a strong background in multivariate statistics and some background in structural equation modeling, equivalent to, for example, the material covered in Tabachnick and Fidell's (2007) book. Conversely, in addition to an adequate background in structural equation modeling, the chapters on multilevel structural equation modeling do not require knowledge of advanced mathematical statistics. In all these cases, I have tried to keep the discussion of the more advanced statistical techniques theoretically sound, but non-technical.

Many of the techniques and their specific software implementations discussed in this book are the subject of active statistical and methodological research. In other words: both the statistical techniques and the software tools are evolving rapidly. As a result, increasing numbers of researchers will apply increasingly advanced models to their data. Of course, researchers still need to understand the models and techniques that they use. Therefore, in addition to being an introduction to multilevel analysis, this book aims to let the reader become acquainted with some advanced modeling techniques that might be used, such as bootstrapping and Bayesian estimation methods. At the time of writing, these are specialist tools, and certainly not part of the standard analysis toolkit. But they are developing rapidly, and are likely to become more popular in applied research as well.

# 2

# The Basic Two-Level Regression Model

The multilevel regression model has become known in the research literature under a variety of names, such as 'random coefficient model' (de Leeuw & Kreft, 1986; Longford, 1993), 'variance component model' (Longford, 1987), and 'hierarchical linear model' (Raudenbush & Bryk, 1986, 1988). Statistically oriented publications tend to refer to the model as a mixed-effects or mixed model (Littell, Milliken, Stroup, & Wolfinger, 1996). The models described in these publications are not *exactly* the same, but they are highly similar, and I will refer to them collectively as 'multilevel regression models'. They all assume that there is a hierarchical data set, with one single outcome or response variable that is measured at the lowest level, and explanatory variables at all existing levels. Conceptually, it is useful to view the multilevel regression model as a hierarchical system of regression equations. In this chapter, I will explain the multilevel regression model for two-level data, and also give an example of three-level data. Regression models with more than two levels are also used in later chapters.

## 2.1 EXAMPLE

Assume that we have data from $J$ classes, with a different number of pupils $n_j$ in each class. On the pupil level, we have the outcome variable 'popularity' ($Y$), measured by a self-rating scale that ranges from 0 (very unpopular) to 10 (very popular). We have two explanatory variables on the pupil level: *pupil gender* ($X_1$: 0 = boy, 1 = girl) and *pupil extraversion* ($X_2$, measured on a self-rating scale ranging from 1 to 10), and one class-level explanatory variable *teacher experience* ($Z$: in years, ranging from 2 to 25). There are data on 2000 pupils in 100 classes, so the average class size is 20 pupils. The data are described in Appendix A.

To analyze these data, we can set up separate regression equations in each class to predict the outcome variable $Y$ using the explanatory variables $X$ as follows:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + e_{ij}. \tag{2.1}$$

Using variable labels instead of algebraic symbols, the equation reads:

$$popularity_{ij} = \beta_{0j} + \beta_{1j}\, gender_{ij} + \beta_{2j}\, extraversion_{ij} + e_{ij}. \tag{2.2}$$

In this regression equation, $\beta_{0j}$ is the intercept, $\beta_{1j}$ is the regression coefficient (regression slope) for the dichotomous explanatory variable gender, $\beta_{2j}$ is the regression coefficient (slope) for the continuous explanatory variable extraversion, and $e_{ij}$ is the usual residual error term. The subscript $j$ is for the classes ($j = 1 \ldots J$) and the subscript $i$ is for individual pupils ($i = 1 \ldots n_j$). The difference with the usual regression model is that we assume that each class has a different intercept coefficient $\beta_{0j}$, and different slope coefficients $\beta_{1j}$ and $\beta_{2j}$. This is indicated in equations 2.1 and 2.2 by attaching a subscript $j$ to the regression coefficients. The residual errors $e_{ij}$ are assumed to have a mean of zero, and a variance to be estimated. Most multilevel software assumes that the variance of the residual errors is the same in all classes. Different authors (see Goldstein, 2003; Raudenbush & Bryk, 2002) use different systems of notation. This book uses $\sigma_e^2$ to denote the variance of the lowest level residual errors.[1]

Since the intercept and slope coefficients are random variables that vary across the classes, they are often referred to as *random* coefficients.[2] In our example, the specific values for the intercept and the slope coefficients are a class characteristic. In general, a class with a high intercept is predicted to have more popular pupils than a class with a low value for the intercept.[3] Similarly, differences in the slope coefficient for gender or extraversion indicate that the relationship between the pupils' gender or extraversion and their predicted popularity is not the same in all classes. Some classes may have a high value for the slope coefficient of gender; in these classes, the difference between boys and girls is relatively large. Other classes may have a low value for the slope coefficient of gender; in these classes, gender has a small effect on the popularity, which means that the difference between boys and girls is small. Variance in the slope for pupil extraversion is interpreted in a similar way; in classes with a large coefficient for the extraversion slope, pupil extraversion has a large impact on their popularity, and vice versa.

Across all classes, the regression coefficients $\beta_j$ are assumed to have a multivariate normal distribution. The next step in the hierarchical regression model is to explain the variation of the regression coefficients $\beta_j$ introducing explanatory variables at the class level:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z_j + u_{0j}, \tag{2.3}$$

[1] At the end of this chapter, a section explains the difference between some commonly used notation systems. Models that are more complicated sometimes need a more complicated notation system, which is introduced in the relevant chapters.

[2] Of course, we hope to explain at least some of the variation by introducing higher-level variables. Generally, we will not be able to explain all the variation, and there will be some unexplained residual variation.

[3] Since the model contains a dummy variable for gender, the precise value of the intercept reflects the predicted value for the boys (coded as zero). Varying intercepts shift the average value for the entire class, both boys and girls.

and

$$\begin{aligned}\beta_{1j} &= \gamma_{10} + \gamma_{11}Z_j + u_{1j} \\ \beta_{2j} &= \gamma_{20} + \gamma_{21}Z_j + u_{2j}.\end{aligned} \tag{2.4}$$

Equation 2.3 predicts the average popularity in a class (the intercept $\beta_{0j}$) by the teacher's experience ($Z$). Thus, if $\gamma_{01}$ is positive, the average popularity is higher in classes with a more experienced teacher. Conversely, if $\gamma_{01}$ is negative, the average popularity is lower in classes with a more experienced teacher. The interpretation of the equations under 2.4 is a bit more complicated. The first equation under 2.4 states that the *relationship*, as expressed by the slope coefficient $\beta_{1j}$, between the popularity ($Y$) and the gender ($X$) of the pupil depends on the amount of experience of the teacher ($Z$). If $\gamma_{11}$ is positive, the gender effect on popularity is larger with experienced teachers. Conversely, if $\gamma_{11}$ is negative, the gender effect on popularity is smaller with experienced teachers. Similarly, the second equation under 2.4 states, if $\gamma_{21}$ is positive, that the effect of extraversion is larger in classes with an experienced teacher. Thus, the amount of experience of the teacher acts as a *moderator variable* for the relationship between popularity and gender or extraversion; this relationship varies according to the value of the moderator variable.

The $u$-terms $u_{0j}$, $u_{1j}$ and $u_{2j}$ in equations 2.3 and 2.4 are (random) residual error terms at the class level. These residual errors $u_j$ are assumed to have a mean of zero, and to be independent from the residual errors $e_{ij}$ at the individual (pupil) level. The variance of the residual errors $u_{0j}$ is specified as $\sigma^2_{u_0}$, and the variance of the residual errors $u_{1j}$ and $u_{2j}$ is specified as $\sigma^2_{u_1}$ and $\sigma^2_{u_2}$. The *covariances* between the residual error terms are denoted by $\sigma_{u_{01}}$, $\sigma_{u_{02}}$ and $\sigma_{u_{12}}$, which are generally *not* assumed to be zero.

Note that in equations 2.3 and 2.4 the regression coefficients $\gamma$ are not assumed to vary across classes. They therefore have no subscript $j$ to indicate to which class they belong. Because they apply to *all* classes, they are referred to as *fixed* coefficients. All between-class variation left in the $\beta$ coefficients, after predicting these with the class variable $Z_j$, is assumed to be residual error variation. This is captured by the residual error terms $u_j$, which do have subscripts $j$ to indicate to which class they belong.

Our model with two pupil-level and one class-level explanatory variable can be written as a single complex regression equation by substituting equations 2.3 and 2.4 into equation 2.1. Rearranging terms gives:

$$\begin{aligned}Y_{ij} = {} & \gamma_{00} + \gamma_{10}X_{1ij} + \gamma_{20}X_{2ij} + \gamma_{01}Z_j + \gamma_{11}X_{1ij}Z_j + \gamma_{21}X_{2ij}Z_j \\ & + u_{1j}X_{1ij} + u_{2j}X_{2ij} + u_{0j} + e_{ij}.\end{aligned} \tag{2.5}$$

Using variable labels instead of algebraic symbols, we have:

$$\begin{aligned} popularity_{ij} = \gamma_{00} + \gamma_{10}\, gender_{ij} + \gamma_{20}\, extraversion_{ij} + \gamma_{01}\, experience_j \\ + \gamma_{11}\, gender_{ij} \times experience_j + \gamma_{21}\, extraversion_{ij} \times experience_j \\ + u_{1j}\, gender_{ij} + u_{2j}\, extraversion_{ij} + u_{0j} + e_{ij}. \end{aligned}$$

The segment $[\gamma_{00} + \gamma_{10}\, X_{1ij} + \gamma_{20}\, X_{2ij} + \gamma_{01} Z_j + \gamma_{11}\, X_{1ij} Z_j + \gamma_{11}\, X_{2ij} Z_j]$ in equation 2.5 contains the fixed coefficients. It is often called the fixed (or deterministic) part of the model. The segment $[u_{1j} X_{1ij} + u_{2j} X_{2ij} + u_{0j} + e_{ij}]$ in equation 2.5 contains the random error terms, and it is often called the random (or stochastic) part of the model. The terms $X_{1i} Z_{jj}$ and $X_{2ij} Z_j$ are interaction terms that appear in the model as a consequence of modeling the varying regression slope $\beta_j$ of a pupil-level variable $X_{ij}$ with the class-level variable $Z_j$. Thus, the moderator effect of $Z$ on the relationship between the dependent variable $Y$ and the predictor $X$ is expressed in the single equation version of the model as a *cross-level interaction*. The interpretation of interaction terms in multiple regression analysis is complex, and this is treated in more detail in Chapter 4. In brief, the point made in Chapter 4 is that the substantive interpretation of the coefficients in models with interactions is much simpler if the variables making up the interaction are expressed as deviations from their respective means.

Note that the random error terms $u_{1j}$ are connected to $X_{ij}$. Since the explanatory variable $X_{ij}$ and the corresponding error term $u_j$ are multiplied, the resulting total error will be different for different values of the explanatory variable $X_{ij}$, a situation that in ordinary multiple regression analysis is called ‘heteroscedasticity’. The usual multiple regression model assumes ‘homoscedasticity’, which means that the variance of the residual errors is independent of the values of the explanatory variables. If this assumption is not true, ordinary multiple regression does not work very well. This is another reason why analyzing multilevel data with ordinary multiple regression techniques does not work well.

As explained in the introduction in Chapter 1, multilevel models are needed because with grouped data observations from the same group are generally more similar to each other than the observations from different groups, and this violates the assumption of independence of all observations. The amount of dependence can be expressed as a correlation coefficient: the intraclass correlation. The methodological literature contains a number of different formulas to estimate the intraclass correlation $\rho$. For example, if we use one-way analysis of variance with the grouping variable as independent variable to test the group effect on our outcome variable, the intraclass correlation is given by $\rho = [MS(B) - MS(error)]/[MS(B)+(n - 1) \times MS(error)]$, where $MS(B)$ is the between groups mean square and $n$ is the common group size. Shrout and Fleiss (1979) give an overview of formulas for the intraclass correlation for a variety of research designs.

If we have simple hierarchical data, the multilevel regression model can also be used to produce an estimate of the intraclass correlation. The model used for this purpose is a model that contains no explanatory variables at all, the so-called *intercept-only* model. The intercept-only model is derived from equations 2.1 and 2.3 as follows. If there are no explanatory variables $X$ at the lowest level, equation 2.1 reduces to:

$$Y_{ij} = \beta_{0j} + e_{ij}. \tag{2.6}$$

Likewise, if there are no explanatory variables $Z$ at the highest level, equation 2.3 reduces to:

$$\beta_{0j} = \gamma_{00} + u_{0j}. \tag{2.7}$$

We find the single equation model by substituting 2.7 into 2.6:

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij}. \tag{2.8}$$

We could also have found equation 2.8 by removing all terms that contain an $X$ or $Z$ variable from equation 2.5. The intercept-only model of equation 2.8 does not explain any variance in $Y$. It only decomposes the variance into two independent components: $\sigma_e^2$, which is the variance of the lowest-level errors $e_{ij}$, and $\sigma_{u0}^2$, which is the variance of the highest-level errors $u_{0j}$. Using this model, we can define the intraclass correlation $\rho$ by the equation:

$$\rho = \frac{\sigma_{u_0}^2}{\sigma_{u_0}^2 + \sigma_e^2}. \tag{2.9}$$

The intraclass correlation $\rho$ indicates the proportion of the variance explained by the grouping structure in the population. Equation 2.9 simply states that the intraclass correlation is the proportion of group-level variance compared to the total variance.[4] The intraclass correlation $\rho$ can also be interpreted as the expected correlation between two randomly drawn units that are in the same group.

Ordinary multiple regression analysis uses an estimation technique called ordinary least squares, abbreviated as OLS. The statistical theory behind the multilevel regression model is more complex, however. Based on observed data, we want to estimate the parameters of the multilevel regression model: the regression coefficients and the variance components. The usual estimators in multilevel regression analysis are

[4] The intraclass correlation is an estimate of the proportion of group-level variance in the *population*. The proportion of group-level variance in the *sample* is given by the correlation ratio $\eta^2$ (eta-squared, see Tabachnick & Fidell, 2007, p. 54): $\eta^2 = SS(B)/SS(Total)$.

maximum likelihood (ML) estimators. Maximum likelihood estimators estimate the parameters of a model by providing estimated values for the population parameters that maximize the so-called 'likelihood function': the function that describes the probability of observing the sample data, given the specific values of the parameter estimates. Simply put, ML estimates are those parameter estimates that maximize the probability of finding the sample data that we have actually found. For an accessible introduction to maximum likelihood methods see Eliason (1993).

Maximum likelihood estimation includes procedures to generate standard errors for most of the parameter estimates. These can be used in significance testing, by computing the test statistic $Z$: $Z = parameter\,/(st.error\,param.)$. This statistic is referred to the standard normal distribution, to establish a $p$-value for the null-hypothesis that the population value of that parameter is zero. The maximum likelihood procedure also produces a statistic called the *deviance*, which indicates how well the model fits the data. In general, models with a lower deviance fit better than models with a higher deviance. If two models are *nested*, meaning that a specific model can be derived from a more general model by removing parameters from that general model, the deviances of the two models can be used to compare their fit statistically. For nested models, the difference in deviance has a chi-square distribution with degrees of freedom equal to the difference in the number of parameters that are estimated in the two models. The deviance test can be used to perform a formal chi-square test, in order to test whether the more general model fits significantly better than the simpler model. The chi-square test of the deviances can also be used to good effect to explore the importance of a set of random effects, by comparing a model that contains these effects against a model that excludes them.

## 2.2 AN EXTENDED EXAMPLE

The intercept-only model is useful as a null model that serves as a benchmark with which other models are compared. For our pupil popularity example data, the intercept-only model is written as:

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij}.$$

The model that includes pupil gender, pupil extraversion and teacher experience, but not the cross-level interactions, is written as:

$$Y_{ij} = \gamma_{00} + \gamma_{10}\,X_{1ij} + \gamma_{20}\,X_{2ij} + \gamma_{01}\,Z_j + u_{1j}\,X_{1ij} + u_{2j}\,X_{2ij} + u_{0j} + e_{ij},$$

or, using variable names instead of algebraic symbols:

$$\begin{aligned} popularity_{ij} = \gamma_{00} + \gamma_{10}\, gender_{ij} + \gamma_{20}\, extraversion_{ij} + \gamma_{01}\, experience_{j} \\ + u_{1j}\, gender_{ij} + u_{2j}\, extraversion_{ij} + u_{0j} + e_{ij}. \end{aligned}$$

*Table 2.1* Intercept-only model and model with explanatory variables

| Model | M0: intercept only | M1: with predictors |
|---|---|---|
| **Fixed part** | Coefficient (s.e.) | Coefficient (s.e.) |
| Intercept | 5.08 (.09) | 0.74 (.20) |
| Pupil gender | | 1.25 (.04) |
| Pupil extraversion | | 0.45 (.03) |
| Teacher experience | | 0.09 (.01) |
| **Random part**[a] | | |
| $\sigma_e^2$ | 1.22 (.04) | 0.55 (.02) |
| $\sigma_{u0}^2$ | 0.69 (.11) | 1.28 (.47) |
| $\sigma_{u1}^2$ | | 0.00 (–) |
| $\sigma_{u2}^2$ | | 0.03 (.008) |
| **Deviance** | 6327.5 | 4812.8 |

[a] For simplicity the covariances are not included.

Table 2.1 presents the parameter estimates and standard errors for both models.[5] In this table, the intercept-only model estimates the intercept as 5.08, which is simply the average popularity across all classes and pupils. The variance of the pupil-level residual errors, symbolized by $\sigma_e^2$, is estimated as 1.22. The variance of the class-level residual errors, symbolized by $\sigma_{u0}^2$, is estimated as 0.69. All parameter estimates are much larger than the corresponding standard errors, and calculation of the *Z*-test shows that they are all significant at $p < .005$.[6] The intraclass correlation, calculated by equation 2.9 as $\rho = \sigma_{u0}^2 \big/ \left(\sigma_{u0}^2 + \sigma_e^2\right)$, is 0.69/1.91, which equals .36. Thus, 36% of the variance of the popularity scores is at the group level, which is very high. Since the intercept-only model contains no explanatory variables, the residual variances represent unexplained error variance. The deviance reported in Table 2.1 is a measure of

[5] For reasons to be explained later, different options for the details of the maximum likelihood procedure may result in slightly different estimates. So, if you re-analyze the example data from this book, the results may differ slightly from the results given here. However, these differences should never be so large that you would draw entirely different conclusions.

[6] Testing variances is preferably done with a test based on the deviance, which is explained in Chapter 3.

model misfit; when we add explanatory variables to the model, the deviance is expected to go down.

The second model in Table 2.1 includes pupil gender and extraversion and teacher experience as explanatory variables. The regression coefficients for all three variables are significant. The regression coefficient for pupil gender is 1.25. Since pupil gender is coded 0 = boy, 1 = girl, this means that on average the girls score 1.25 points higher on the popularity measure. The regression coefficient for pupil extraversion is 0.45, which means that with each scale point higher on the extraversion measure, the popularity is expected to increase by 0.45 scale points. The regression coefficient for teacher experience is 0.09, which means that for each year of experience of the teacher, the average popularity score of the class goes up by 0.09 points. This does not seem very much, but the teacher experience in our example data ranges from 2 to 25 years, so the predicted difference between the least experienced and the most experienced teacher is $(25 - 2 =)\ 23 \times 0.09 = 2.07$ points on the popularity measure. We can use the standard errors of the regression coefficients reported in Table 2.1 to construct a 95% confidence interval. For the regression coefficient of pupil gender, the 95% confidence interval runs from 1.17 to 1.33, the confidence interval for pupil extraversion runs from 0.39 to 0.51, and the 95% confidence interval for the regression coefficient of teacher experience runs from 0.07 to 0.11.

The model with the explanatory variables includes variance components for the regression coefficients of pupil gender and pupil extraversion, symbolized by $\sigma_{u1}^2$ and $\sigma_{u2}^2$ in Table 2.1. The variance of the regression coefficients for pupil extraversion across classes is estimated as 0.03, with a standard error of .008. The simple $Z$-test ($Z = 3.75$) results in a (one sided) $p$-value of $p < .001$, which is clearly significant. The variance of the regression coefficients for pupil gender is estimated as 0.00. This variance component is clearly not significant, so the hypothesis that the regression slopes for pupil gender vary across classes is not supported by the data. Therefore we can remove the residual variance term for the gender slopes from the model.[7] Table 2.2 presents the estimates for the model with a fixed slope for the effect of pupil gender. Table 2.2 also includes the covariance between the class-level errors for the intercept and the extraversion slope. These covariances are rarely interpreted, and for that reason often not included in the reported tables. However, as Table 2.2 demonstrates, they can be quite large and significant, so as a rule they are always included in the model.

The significant variance of the regression slopes for pupil extraversion implies that we should not interpret the estimated value of 0.45 without considering this

[7] Multilevel software deals with the problem of zero variances in different ways. Most software inserts a zero which may or may not be flagged as a redundant parameter. In general, such zero variances should be removed from the model, and the resulting new model must be re-estimated.

*Table 2.2* Model with explanatory variables, extraversion slope random

| Model | M1: with predictors |
|---|---|
| **Fixed part** | Coefficient (s.e.) |
| Intercept | 0.74 (.20) |
| Pupil gender | 1.25 (.04) |
| Pupil extraversion | 0.45 (.02) |
| Teacher experience | 0.09 (.01) |
| **Random part** | |
| $\sigma_e^2$ | 0.55 (.02) |
| $\sigma_{u0}^2$ | 1.28 (.28) |
| $\sigma_{u2}^2$ | 0.03 (.008) |
| $\sigma_{u_{02}}$ | –0.18 (.05) |
| **Deviance** | 4812.8 |

variation. In an ordinary regression model, without multilevel structure, the value of 0.45 means that for each point of difference on the extraversion scale, pupil popularity goes up by 0.45, for all pupils in all classes. In our multilevel model, the regression coefficient for pupil gender varies across the classes, and the value of 0.45 is just the expected value (the mean) across all classes. The varying regression slopes for pupil extraversion are assumed to follow a normal distribution. The variance of this distribution is in our example estimated as 0.034. Interpretation of this variation is easier when we consider the standard deviation, which is the square root of the variance or 0.18 in our example data. A useful characteristic of the standard deviation is that with normally distributed observations about 67% of the observations lie between one standard deviation below and one above the mean, and about 95% of the observations lie between two standard deviations below and above the mean. If we apply this to the regression coefficients for pupil gender, we conclude that about 67% of the regression coefficients are expected to lie between (0.45 – 0.18 =) 0.27 and (0.45 + 0.18 =) 0.63, and about 95% are expected to lie between (0.45 – 0.37 =) 0.08 and (0.45 + 0.37 =) 0.82. The more precise value of $Z_{.975} = 1.96$ leads to the 95% predictive interval calculated as 0.09 to 0.81. We can also use the standard normal distribution to estimate the percentage of regression coefficients that are negative. As it turns out, if the mean regression coefficient for pupil extraversion is 0.45, given the estimated slope variance, less than 1% of the classes are expected to have a regression coefficient that is actually negative. Note that the 95% interval computed here is totally different from the 95% confidence interval for the regression coefficient of pupil extraversion, which runs from 0.41 to 0.50. The 95% confidence interval applies to $\gamma_{20}$, the mean value of the regression coefficients across all the classes. The 95% interval calculated here is the 95% *predictive interval*, which expresses that 95% of the regression coefficients of the variable 'pupil extraversion' in the classes are predicted to lie between 0.09 and 0.81.

Given the significant variance of the regression coefficient of pupil extraversion across the classes it is attractive to attempt to predict its variation using class-level variables. We have one class-level variable: teacher experience. The individual-level regression equation for this example, using variable labels instead of symbols, is given by equation 2.10:

$$popularity_{ij} = \beta_{0j} + \beta_1\ gender_{ij} + \beta_{2j}\ extraversion_{ij} + e_{ij}. \tag{2.10}$$

The regression coefficient $\beta_1$ for pupil gender does not have a subscript $j$, because it is not assumed to vary across classes. The regression equations predicting $\beta_{0j}$, the intercept in class $j$, and $\beta_{2j}$, the regression slope of pupil extraversion in class $j$, are given by equation 2.3 and 2.4, which are rewritten below using variable labels:

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01} t.exp_j + u_{0j}\\ \beta_{2j} &= \gamma_{20} + \gamma_{21} t.exp_j + u_{2j}.\end{aligned} \tag{2.11}$$

By substituting 2.11 into 2.10 we get:

$$\begin{aligned}popularity_{ij} = {}& \gamma_{00} + \gamma_{10}\ gender_{ij} + \gamma_{20}\ extraversion_{ij} + \gamma_{01} t.exp_j\\ & + \gamma_{21} extraversion_{ij}\ t.exp_j + u_{2j}\ extraversion_{ij} + u_{0j} + e_{ij}\end{aligned} \tag{2.12}$$

The algebraic manipulations of the equations above make clear that to explain the variance of the regression slopes $\beta_{2j}$, we need to introduce an interaction term in the model. This interaction, between the variables pupil extraversion and teacher experience, is a cross-level interaction, because it involves explanatory variables from different levels. Table 2.3 presents the estimates from a model with this cross-level interaction. For comparison, the estimates for the model without this interaction are also included in Table 2.3.

The estimates for the fixed coefficients in Table 2.3 are similar for the effect of pupil gender, but the regression slopes for pupil extraversion and teacher experience are considerably larger in the cross-level model. The interpretation remains the same: extraverted pupils are more popular. The regression coefficient for the cross-level interaction is –0.03, which is small but significant. This interaction is formed by multiplying the scores for the variables 'pupil extraversion' and 'teacher experience', and the negative value means that with experienced teachers, the advantage of being

*Table 2.3* Model without and with cross-level interaction

| Model | M1A: main effects | M2: with interaction |
|---|---|---|
| **Fixed part** | Coefficient (s.e.) | Coefficient (s.e.) |
| Intercept | 0.74 (.20) | –1.21 (.27) |
| Pupil gender | 1.25 (.04) | 1.24 (.04) |
| Pupil extraversion | 0.45 (.02) | 0.80 (.04) |
| Teacher experience | 0.09 (.01) | 0.23 (.02) |
| Extra × T.exp | | –0.03 (.003) |
| **Random part** | | |
| $\sigma_e^2$ | 0.55 (.02) | 0.55 (.02) |
| $\sigma_{u0}^2$ | 1.28 (.28) | 0.45 (.16) |
| $\sigma_{u2}^2$ | 0.03 (.008) | 0.005 (.004) |
| $\sigma_{u_{02}}$ | –0.18 (.05) | –0.03 (.02) |
| **Deviance** | 4812.8 | 4747.6 |

extraverted is smaller than expected from the direct effects only. Thus, the difference between extraverted and introverted pupils is smaller with more experienced teachers.

Comparison of the other results between the two models shows that the variance component for pupil extraversion goes down from 0.03 in the direct effects model to 0.005 in the cross-level model. Apparently, the cross-level model explains some of the variation of the slopes for pupil extraversion. The deviance also goes down, which indicates that the model fits better than the previous model. The other differences in the random part are more difficult to interpret. Much of the difficulty in reconciling the estimates in the two models in Table 2.3 stems from adding an interaction effect between variables that have not been centered. This issue is discussed in more detail in Chapter 4.

The coefficients in the tables are all unstandardized regression coefficients. To interpret them properly, we must take the scale of the explanatory variables into account. In multiple regression analysis, and structural equation models, for that matter, the regression coefficients are often standardized because that facilitates the interpretation when one wants to compare the effects of different variables within one sample. Only if the goal of the analysis is to compare parameter estimates from different samples to each other, should one always use unstandardized coefficients. To standardize the regression coefficients, as presented in Table 2.1 or Table 2.3, one could standardize all variables before putting them into the multilevel analysis. However, this would in general also change the estimates of the variance components. This may not be a bad thing in itself, because standardized variables are also centered on their overall mean. Centering explanatory variables has some distinct advantages, which are discussed in Chapter 4. Even so, it is also possible to derive the standardized regression coefficients from the unstandardized coefficients:

$$standardized\ coefficient = \frac{unstandardized\ coefficient \times stand.dev.explanatory\ var.}{stand.dev.outcome\ var.} \tag{2.13}$$

In our example data, the standard deviations are: 1.38 for popularity, 0.51 for gender, 1.26 for extraversion, and 6.55 for teacher experience. Table 2.4 presents the unstandardized and standardized coefficients for the second model in Table 2.2. It also presents the estimates that we obtain if we first standardize all variables, and then carry out the analysis.

*Table 2.4* Comparing unstandardized and standardized estimates

| Model | Standardization using 2.13 | | Standardized variables |
|---|---|---|---|
| **Fixed part** | Coefficient (s.e.) | Standardized | Coefficient (s.e.) |
| Intercept | 0.74 (.20) | – | –0.03 (.04) |
| Pupil gender | 1.25 (.04) | 0.46 | 0.45 (.01) |
| Pupil extraversion | 0.45 (.02) | 0.41 | 0.41 (.02) |
| Teacher experience | 0.09 (.01) | 0.43 | 0.43 (.04) |
| **Random part** | | | |
| $\sigma_e^2$ | 0.55 (.02) | | 0.28 (.01) |
| $\sigma_{u0}^2$ | 1.28 (.28) | | 0.15 (.02) |
| $\sigma_{u2}^2$ | 0.03 (.008) | | 0.03 (.01) |
| $\sigma_{u_{02}}$ | –0.18 (.05) | | –0.01 (.01) |
| **Deviance** | 4812.8 | | 3517.2 |

Table 2.4 shows that the standardized regression coefficients are almost the same as the coefficients estimated for standardized variables. The small differences in Table 2.4 are simply a result of rounding errors. However, if we use standardized variables in our analysis, we find very different variance components and a very different value for the deviance. This is not only the effect of scaling the variables differently, which becomes clear if we realize that the covariance between the slope for pupil extraversion and the intercept is significant for the unstandardized variables, but not significant for

the standardized variables. This kind of difference in results is general. The fixed part of the multilevel regression model is invariant for linear transformations, just like the regression coefficients in the ordinary single-level regression model. This means that if we change the scale of our explanatory variables, the regression coefficients and the corresponding standard errors change by the same multiplication factor, and all associated *p*-values remain exactly the same. However, the random part of the multilevel regression model is not invariant for linear transformations. The estimates of the variance components in the random part can and do change, sometimes dramatically. This is discussed in more detail in section 4.2 in Chapter 4. The conclusion to be drawn here is that, if we have a complicated random part, including random components for regression slopes, we should think carefully about the scale of our explanatory variables. If our only goal is to present standardized coefficients in addition to the unstandardized coefficients, applying equation 2.13 is safer than transforming our variables. On the other hand, we may estimate the unstandardized results, including the random part and the deviance, and then re-analyze the data using standardized variables, merely using this analysis as a computational trick to obtain the standardized regression coefficients without having to do hand calculations.

## 2.3 INSPECTING RESIDUALS

Inspection of residuals is a standard tool in multiple regression analysis to examine whether assumptions of normality and linearity are met (see Stevens, 2009; Tabachnick & Fidell, 2007). Multilevel regression analysis also assumes normality and linearity. Since the multilevel regression model is more complicated than the ordinary regression model, checking such assumptions is even more important. For example, Bauer and Cai (2009) show that neglecting a nonlinear relationship may result in spuriously high estimates of slope variances and cross-level interaction effects. Inspection of the residuals is one way to investigate linearity and homoscedasticity. There is one important difference from ordinary regression analysis; we have more than one residual, in fact, we have residuals for each random effect in the model. Consequently, many different residuals plots can be made.

### 2.3.1 Examples of residuals plots

The equation below represents the one-equation version of the direct effects model for our example data. This is the multilevel model without the cross-level interaction. Since the interaction explains part of the extraversion slope variance, a model that does not include this interaction produces a graph that displays the actual slope variation more fully.

$$\begin{aligned} popularity_{ij} = {} & \gamma_{00} + \gamma_{10}\, gender_{ij} + \gamma_{20}\, extraversion_{ij} + \gamma_{01}\, experience_j \\ & + u_{2j}\, extraversion_{ij} + u_{0j} + e_{ij} \end{aligned}$$

In this model, we have three residual error terms: $e_{ij}$, $u_{0j}$, and $u_{2j}$. The $e_{ij}$ are the residual prediction errors at the lowest level, similar to the prediction errors in ordinary single-level multiple regression. A simple boxplot of these residuals will enable us to identify extreme outliers. An assumption that is usually made in multilevel regression analysis is that the variance of the residual errors is the same in all groups. This can be assessed by computing a one-way analysis of variance of the groups on the absolute values of the residuals, which is the equivalent of Levene's test for equality of variances in analysis of variance (Stevens, 2009). Raudenbush and Bryk (2002) describe a chi-square test that can be used for the same purpose.

The $u_{0j}$ are the residual prediction errors at the group level, which can be used in ways analogous to the investigation of the lowest-level residuals $e_{ij}$. The $u_{2j}$ are the residuals of the regression slopes across the groups. By plotting the regression slopes for the various groups, we get a visual impression of how much the regression slopes actually differ, and we may also be able to identify groups which have a regression slope that is wildly different from the others.

To test the normality assumption, we can plot standardized residuals against their normal scores. If the residuals have a normal distribution, the plot should show a straight diagonal line. Figure 2.1 is a scatterplot of the standardized level 1 residuals, calculated for the final model including cross-level interaction, against their normal scores. The graph indicates close conformity to normality, and no extreme outliers. Similar plots can be made for the level 2 residuals.

We obtain a different plot, if we plot the residuals against the predicted values of the outcome variable popularity, using the fixed part of the multilevel regression model for the prediction. Such a scatter plot of the residuals against the predicted values provides information about possible failure of normality, nonlinearity, and heteroscedasticity. If these assumptions are met, the plotted points should be evenly divided above and below their mean value of zero, with no strong structure (see Tabachnick & Fidell, 2007, p. 162). Figure 2.2 shows this scatter plot for the level 1 residuals. For our example data, the scatter plot in Figure 2.2 does not indicate strong violations of the assumptions.

Similar scatter plots can be made for the second-level residuals for the intercept and the slope of the explanatory variable pupil extraversion. As an illustration, Figure 2.3 shows the scatterplots of the level 2 residuals around the average intercept and around the average slope of pupil extraversion against the predicted values of the outcome variable popularity. Both scatterplots indicate that the assumptions are reasonably met.

An interesting plot that can be made using the level 2 residuals is a plot of the

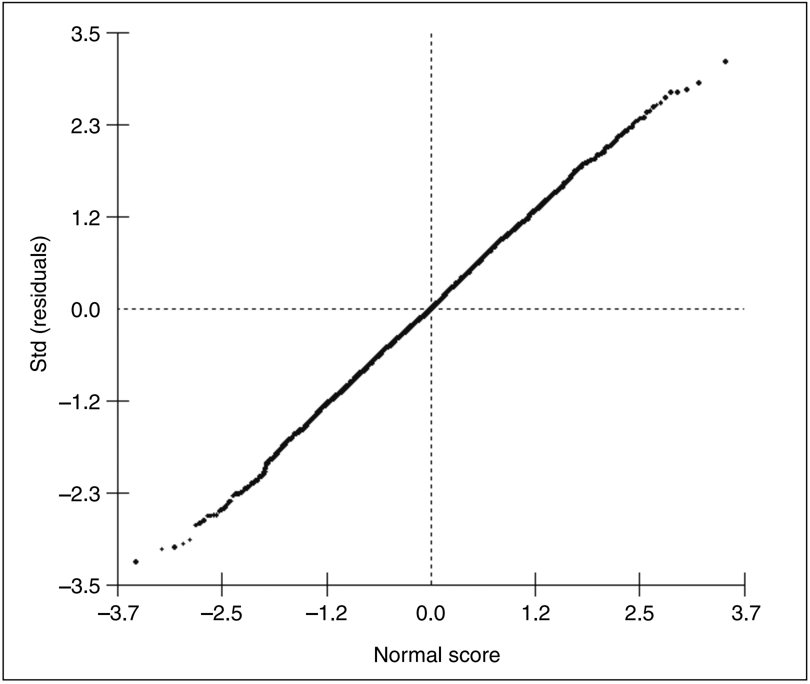


*Figure 2.1* Plot of level 1 standardized residuals against normal scores.

residuals against their rank order, with an added error bar. In Figure 2.4, an error bar frames each point estimate, and the classes are sorted in rank order of the residuals. The error bars represent the confidence interval around each estimate, constructed by multiplying its standard error by 1.39 instead of the more usual 1.96. Using 1.39 as the multiplication factor results in confidence intervals with the property that if the error bars of two classes do not overlap, they have significantly different residuals at the 5% level (Goldstein, 2003). For a discussion of the construction and use of these error bars see Goldstein and Healy (1995) and Goldstein and Spiegelhalter (1996). In our example, this plot, sometimes called the *caterpillar* plot, shows some outliers at each end. This gives an indication of exceptional residuals for the intercept. A logical next step would be to identify the classes at the extremes of the rank order, and to seek a post hoc interpretation of what makes these classes different.

Examining residuals in multivariate models presents us with a problem. For instance, the residuals should show a nice normal distribution, which implies an absence of extreme outliers. However, this applies to the residuals after including all important explanatory variables and relevant parameters in the model. If we analyze a

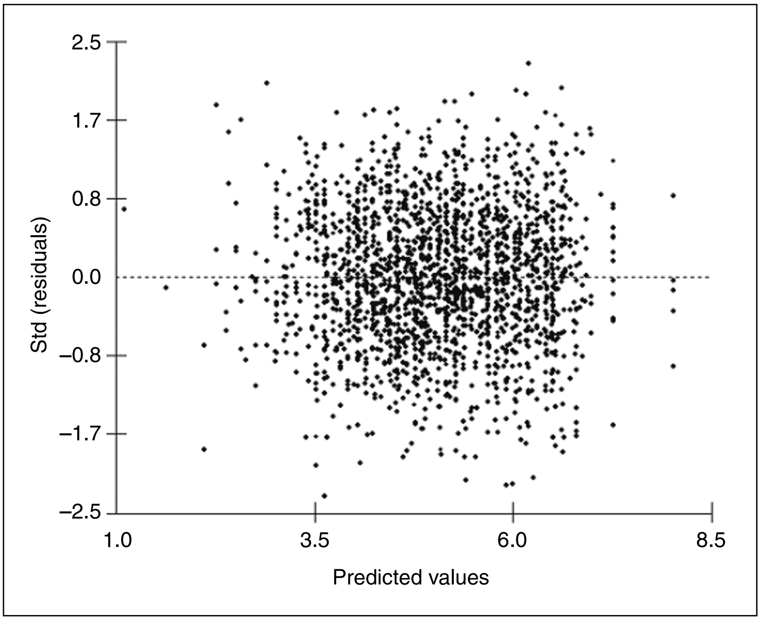


*Figure 2.2* Level 1 standardized residuals plotted against predicted popularity.

sequence of models, we have a series of different residuals for each model, and scrutinizing them all at each step is not always practical. On the other hand, our decision to include a specific variable or parameter in our model might well be influenced by a violation of some assumption. Although there is no perfect solution to this dilemma, a reasonable approach is to examine the two residual terms in the intercept-only model, to find out if there are gross violations of the assumptions of the model. If there are, we should accommodate them, for instance by applying a normalizing transformation, by deleting certain individuals or groups from our data set, or by including a dummy variable that indicates a specific outlying individual or group. When we have determined our final model, we should make a more thorough examination of the various residuals. If we detect gross violations of assumptions, these should again be accommodated, and the model should be estimated again. Of course, after accommodating an extreme outlier, we might find that a previously significant effect has disappeared, and that we need to change our model again. Procedures for model exploration and detection of violations in ordinary multiple regression are discussed, for instance, in Tabachnick and Fidell (2007) or Field (2009). In multilevel regression, the same procedures apply, but the analyses are more complicated because we have to examine more than one set of residuals, and must distinguish between multiple levels.

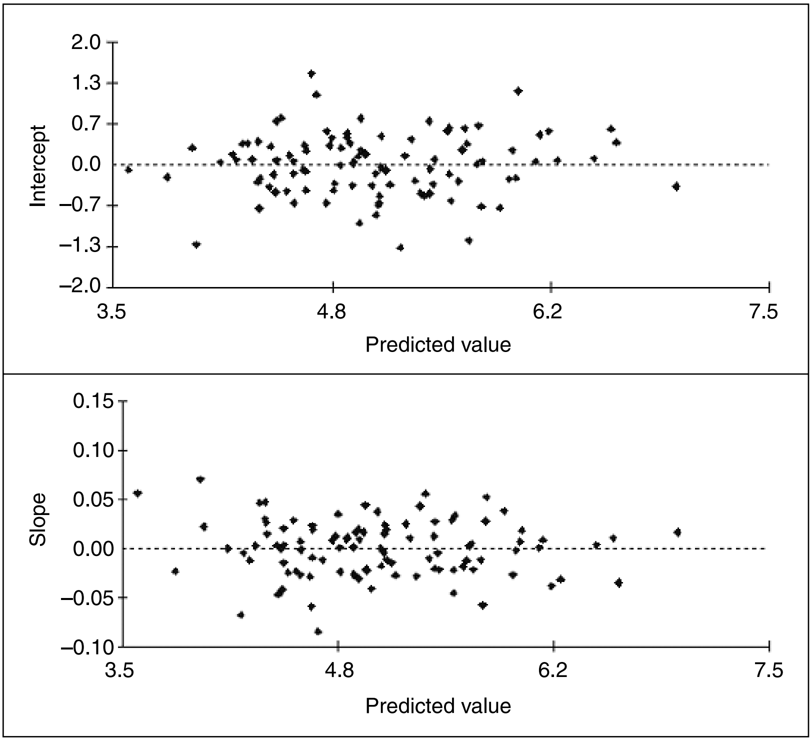


*Figure 2.3* Level 2 residuals plotted against predicted popularity.

As mentioned at the beginning of this section, graphs can be useful in detecting outliers and nonlinear relations. However, an observation may have an undue effect on the outcome of a regression analysis without being an obvious outlier. Figure 2.5, a scatter plot of the so-called Anscombe data (Anscombe, 1973), illustrates this point. There is one data point in Figure 2.5, which by itself almost totally determines the regression line. Without this one observation, the regression line would be very different. Yet, when the residuals are inspected, it does not show up as an obvious outlier.

In ordinary regression analysis, various measures have been proposed to indicate the influence of individual observations on the outcome (see Tabachnick & Fidell, 2007). In general, such *influence* or *leverage* measures are based on a comparison of the estimates when a specific observation is included in the data or not. Langford and Lewis (1998) discuss extensions of these influence measures for the multilevel regression model. Since most of these measures are based on comparison of estimates with and without a specific observation, it is difficult to calculate them by hand. However, if

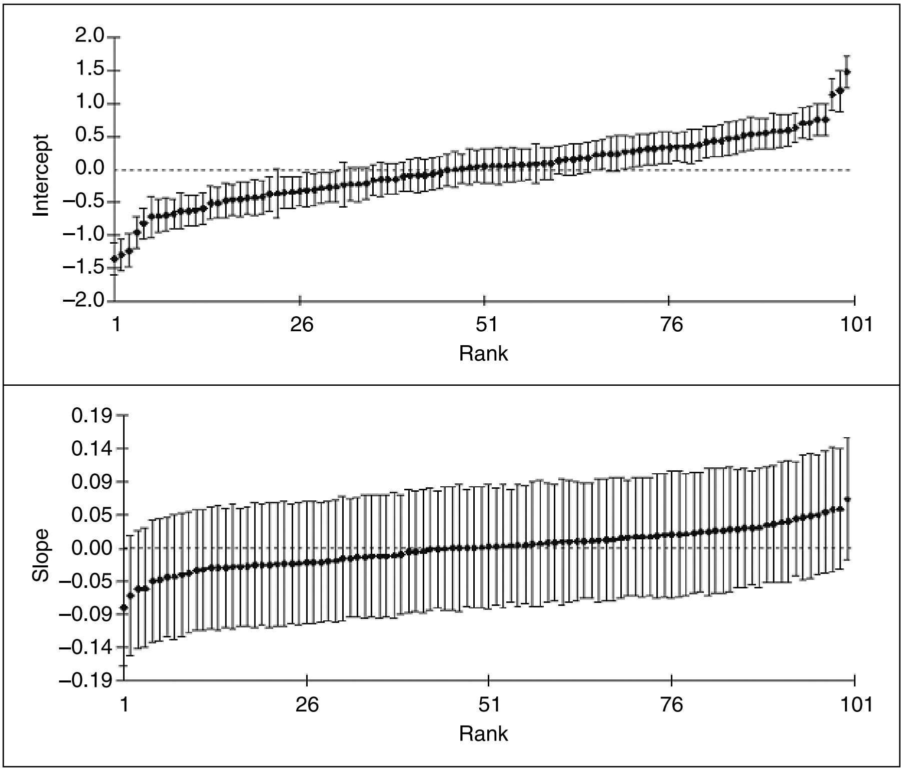


*Figure 2.4* Error bar plot of level 2 residuals.

the software offers the option to calculate influence measures, it is advisable to do so. If a unit (individual or group) has a large value for the influence measure, that specific unit has a large influence on the values of the regression coefficients. It is useful to inspect cases with extreme influence values for possible violations of assumptions, or even data errors.

### 2.3.2 Examining slope variation: OLS and shrinkage estimators

The residuals can be added to the average values of the intercept and slope, to produce predictions of the intercepts and slopes in different groups. These can also be plotted.

For example, Figure 2.6 plots the 100 regression slopes for the explanatory variable pupil extraversion in the 100 classes. It is clear that for most classes the effect is strongly positive: extravert pupils tend to be more popular in all classes. It is also clear that in some classes the relationship is more pronounced than in other classes. Most of

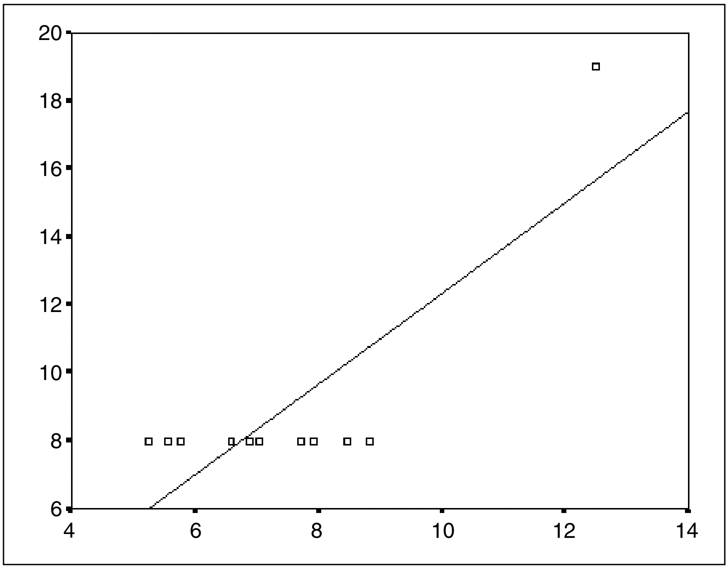


*Figure 2.5* Regression line determined by one single observation.

the regression slopes are not very different from the others, although there are a few slopes that are clearly different. It could be useful to examine the data for these classes in more detail, to find out if there is a reason for the unusual slopes.

The predicted intercepts and slopes for the 100 classes are not identical to the values we would obtain if we carried out 100 separate ordinary regression analyses in each of the 100 classes, using standard ordinary least squares (OLS) techniques. If we were to compare the results from 100 separate OLS regression analyses to the values obtained from a multilevel regression analysis, we would find that the results from the separate analyses are more variable. This is because the multilevel estimates of the regression coefficients of the 100 classes are weighted. They are so-called Empirical Bayes (EB) or *shrinkage* estimates: a weighted average of the specific OLS estimate in each class and the overall regression coefficient, estimated for all similar classes.

As a result, the regression coefficients are *shrunk* back towards the mean coefficient for the whole data set. The shrinkage weight depends on the reliability of the estimated coefficient. Coefficients that are estimated with small accuracy shrink more than very accurately estimated coefficients. Accuracy of estimation depends on two factors: the group sample size, and the distance between the group-based estimate and the overall estimate. Estimates for small groups are less reliable, and shrink more than estimates for large groups. Other things being equal, estimates that are very far from the overall estimate are assumed less reliable, and they shrink more than estimates that

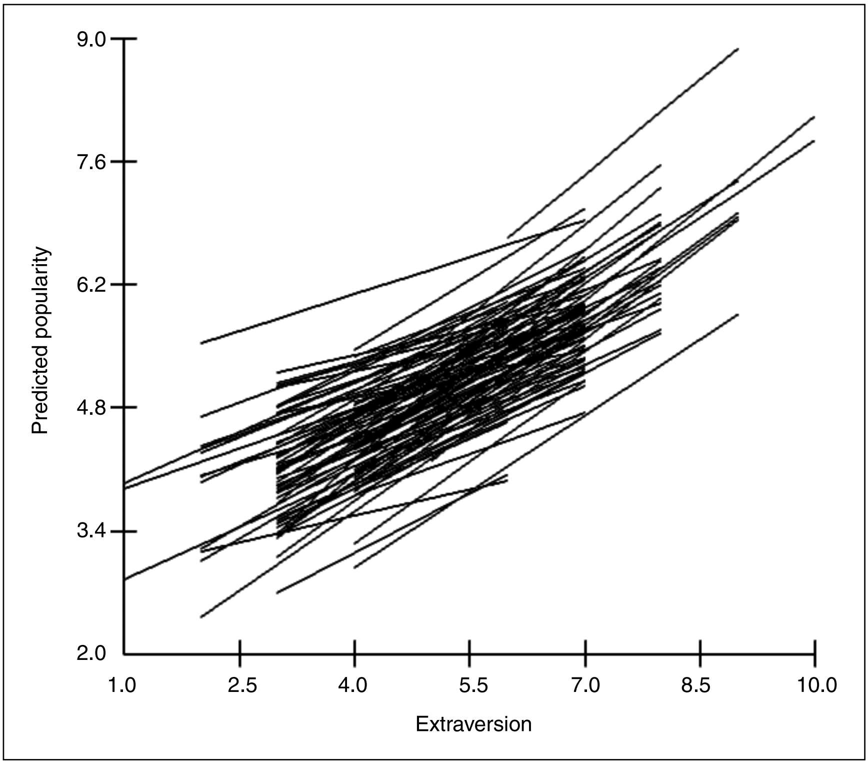


*Figure 2.6* Plot of the 100 class regression slopes for pupil extraversion.

are close to the overall average. The statistical method used is called *empirical Bayes estimation*. Because of this shrinkage effect, empirical Bayes estimators are biased. However, they are usually more precise, a property that is often more useful than being unbiased (see Kendall, 1959).

The equation to form the empirical Bayes estimate of the intercepts is given by:

$$\hat{\beta}_{0j}^{\mathrm{EB}} = \lambda_j \hat{\beta}_{0j}^{\mathrm{OLS}} + (1 - \lambda_j)\gamma_{00}, \tag{2.14}$$

where $\lambda_j$ is the reliability of the OLS estimate $\beta_{0j}^{\mathrm{OLS}}$ as an estimate of $\beta_{0j}$, which is given by the equation $\lambda_j = \sigma_{u_0}^2 / \left(\sigma_{u_0}^2 + \sigma_e^2 / n_j\right)$ (Raudenbush & Bryk, 2002), and $\gamma_{00}$ is the overall intercept. The reliability $\lambda_j$ is close to 1.0 when the group sizes are large and/or the variability of the intercepts across groups is large. In these cases, the overall estimate $\gamma_{00}$ is not a good indicator of each group's intercept. If the group sizes are small and there is little variation across groups, the reliability $\lambda_j$ is close to 0.0, and more weight is put on the overall estimate $\gamma_{00}$. Equation 2.14 makes clear that, since the OLS estimates are unbiased, the empirical Bayes estimates $\beta_{0j}^{\mathrm{EB}}$ must be biased towards the overall estimate $\gamma_{00}$. They are *shrunken* towards the average value $\gamma_{00}$. For that reason, the empirical Bayes estimators are also referred to as shrinkage estimators. Figure 2.7 presents boxplots for the OLS and EB estimates of the intercept and the extraversion regression slopes in the model without the cross-level interaction (model M1A in Table 2.3). It is clear that the OLS estimates have a higher variability.

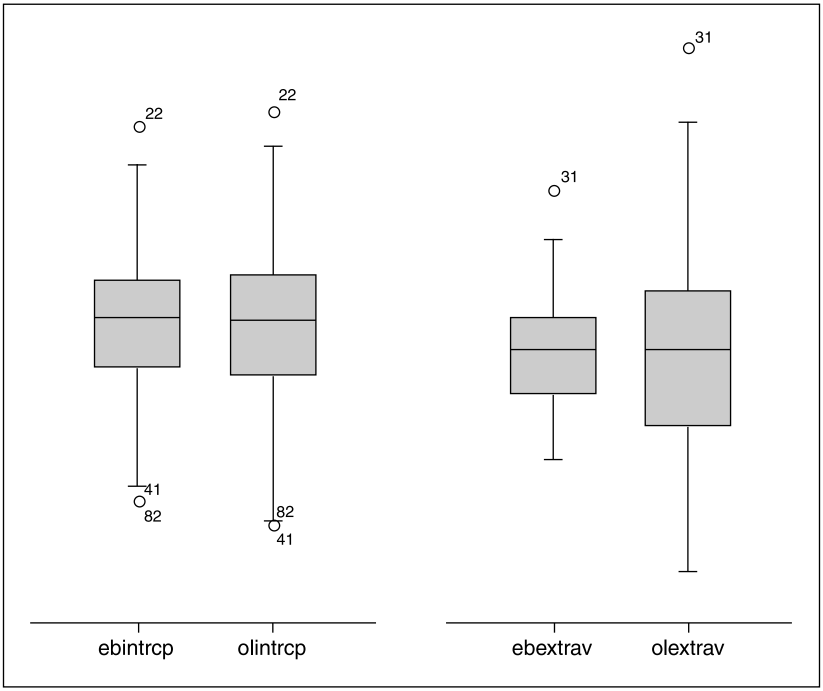


*Figure 2.7* OLS and EB estimates for intercept and slope.

Although the empirical Bayes or shrinkage estimators are biased, they are also in general closer to the (unknown) values of $\beta_{0j}$ (Bryk & Raudenbush, 1992, p. 40). If the regression model includes a group-level model, the shrinkage estimators are conditional on the group-level model. The advantages of shrinkage estimators remain, provided that the group-level model is well specified (Bryk & Raudenbush, 1992, p. 80). This is especially important if the estimated coefficients are used to describe specific groups. For instance, we can use estimates for the intercepts of the schools to rank them on their average outcome. If this is used as an indicator of the quality of schools, the shrinkage estimators introduce a bias, because high scoring schools will be presented too negatively, and low scoring schools will be presented too positively. This is offset by the advantage of having a smaller standard error (Carlin & Louis, 1996; Lindley & Smith, 1972). Bryk and Raudenbush discuss this problem in an example involving the effectiveness of organizations (Bryk & Raudenbush, 1992, Chapter 5); see also the cautionary points made by Raudenbush and Willms (1991) and Snijders and Bosker (1999, pp. 58–63). All emphasize that the higher precision of the empirical Bayes residuals is achieved at the expense of a certain bias. The bias is largest when we inspect groups that are both small and far removed from the overall mean. In such cases, inspecting residuals should be supplemented with other procedures, such as comparing error bars for all schools (Goldstein & Healy, 1995). Error bars are illustrated in this chapter in Figure 2.4.

## 2.4 THREE- AND MORE-LEVEL REGRESSION MODELS

### 2.4.1 Multiple-level models

In principle, the extension of the two-level regression model to three and more levels is straightforward. There is an outcome variable at the first, the lowest level. In addition, there may be explanatory variables at all available levels. The problem is that three- and more-level models can become complicated very fast. In addition to the usual fixed regression coefficients, we must entertain the possibility that regression coefficients for first-level explanatory variables may vary across units of both the second and the third level. Regression coefficients for second-level explanatory variables may vary across units of the third level. To explain such variation, we must include cross-level interactions in the model. Regression slopes for the cross-level interaction between first-level and second-level variables may themselves vary across third-level units. To explain such variation, we need a three-way interaction involving variables at all three levels.

The equations for such models are complicated, especially when we do not use the more compact summation notation but write out the complete single-equation version of the model in an algebraic format (for a note on notation see section 2.5).

The resulting models are not only difficult to follow from a conceptual point of view; they may also be difficult to estimate in practice. The number of estimated parameters is considerable, and at the same time the highest-level sample size tends to become relatively smaller. As DiPrete and Forristal (1994, p. 349) put it, the imagination of the researchers '. . . can easily outrun the capacity of the data, the computer, and current optimization techniques to provide robust estimates.'

Nevertheless, three- and more-level models have their place in multilevel analysis. Intuitively, three-level structures such as pupils in classes in schools, or respondents nested within households, nested within regions, appear to be both conceptually and empirically manageable. If the lowest level is repeated measures over time, having repeated measures on pupils nested within schools again does not appear to be overly complicated. In such cases, the solution for the conceptual and statistical problems mentioned is to keep models reasonably small. Especially specification of the higher-level variances and covariances should be driven by theoretical considerations. A higher-level variance for a specific regression coefficient implies that this regression coefficient is assumed to vary across units at that level. A higher-level covariance between two specific regression coefficients implies that these regression coefficients are assumed to covary across units at that level. Especially when models become large and complicated, it is advisable to avoid higher-order interactions, and to include in the random part only those elements for which there is strong theoretical or empirical justification. This implies that an exhaustive search for second-order and higher-order interactions is not a good idea. In general, we should look for higher-order interactions only if there is strong theoretical justification for their importance, or if an unusually large variance component for a regression slope calls for explanation. For the random part of the model, there are usually more convincing theoretical reasons for the higher-level variance components than for the covariance components. Especially if the covariances are small and insignificant, analysts sometimes do not include all possible covariances in the model. This is defensible, with some exceptions. First, it is recommended that the covariances between the intercept and the random slopes are always included. Second, it is recommended to include covariances corresponding to slopes of dummy variables belonging to the same categorical variable, and for variables that are involved in an interaction or belong to the same polynomial expression (Longford, 1990, pp. 79–80).

### 2.4.2 Intraclass correlations in three-level models

In a two-level model, the intraclass correlation is calculated in the intercept-only model using equation 2.9, which is repeated below:

$$\rho = \frac{\sigma_{u_0}^2}{\sigma_{u_0}^2 + \sigma_e^2}. \qquad \text{(2.9, repeated)}$$

The intraclass correlation is an indication of the proportion of variance at the second level, and it can also be interpreted as the expected (population) correlation between two randomly chosen individuals within the same group.

If we have a three-level model, for instance pupils nested within classes, nested within schools, there are several ways to calculate the intraclass correlation. First, we estimate an intercept-only model for the three-level data, for which the single-equation model can be written as follows:

$$Y_{ijk} = \gamma_{000} + v_{0k} + u_{0jk} + e_{ijk}. \tag{2.15}$$

The variances at the first, second, and third level are respectively $\sigma_e^2$, $\sigma_{u_0}^2$, and $\sigma_{v_0}^2$. The first method (see Davis & Scott, 1995) defines the intraclass correlations at the class and school level as:

$$\rho_{class} = \frac{\sigma_{u_0}^2}{\sigma_{v_0}^2 + \sigma_{u_0}^2 + \sigma_e^2}, \tag{2.16}$$

and:

$$\rho_{school} = \frac{\sigma_{v_0}^2}{\sigma_{v_0}^2 + \sigma_{u_0}^2 + \sigma_e^2}. \tag{2.17}$$

The second method (see Siddiqui, Hedeker, Flay, & Hu, 1996) defines the intraclass correlations at the class and school level as:

$$\rho_{class} = \frac{\sigma_{v_0}^2 + \sigma_{u_0}^2}{\sigma_{v_0}^2 + \sigma_{u_0}^2 + \sigma_e^2}, \tag{2.18}$$

and:

$$\rho_{school} = \frac{\sigma_{v_0}^2}{\sigma_{v_0}^2 + \sigma_{u_0}^2 + \sigma_e^2}. \tag{2.19}$$

Actually, both methods are correct (Algina, 2000). The first method identifies the proportion of variance at the class and school level. This should be used if we are interested in a decomposition of the variance across the available levels, or if we are interested in how much variance is explained at each level (a topic discussed in section 4.5). The second method represents an estimate of the expected (population) correlation between two randomly chosen elements in the same group. So $\rho_{class}$ as calculated in equation 2.18 is the expected correlation between two pupils within the same class, and it correctly takes into account that two pupils who are in the same class must by definition also be in the same school. For this reason, the variance components for classes and schools must both be in the numerator of equation 2.18. If the two sets of estimates are different, which may happen if the amount of variance at the school level is large, there is no contradiction involved. Both sets of equations express two different aspects of the data, which happen to coincide when there are only two levels.

### 2.4.3 An example of a three-level model

The data in this example are from a hypothetical study on stress in hospitals. The data are from nurses working in wards nested within hospitals. In each of 25 hospitals, four wards are selected and randomly assigned to an experimental and control condition. In the experimental condition, a training program is offered to all nurses to cope with job-related stress. After the program is completed, a sample of about 10 nurses from each ward is given a test that measures job-related stress. Additional variables are: nurse age (years), nurse experience (years), nurse gender (0 = male, 1 = female), type of ward (0 = general care, 1 = special care), and hospital size (0 = small, 1 = medium, 2 = large).

This is an example of an experiment where the experimental intervention is carried out at the group level. In biomedical research this design is known as a cluster randomized trial. They are quite common also in educational and organizational research, where entire classes or schools are assigned to experimental and control conditions. Since the design variable Experimental versus Control group (*ExpCon*) is manipulated at the second (ward) level, we can study whether the experimental effect is different in different hospitals, by defining the regression coefficient for the *ExpCon* variable as random at the hospital level.

In this example, the variable *ExpCon* is of main interest, and the other variables are covariates. Their function is to control for differences between the groups, which should be small given that randomization is used, and to explain variance in the outcome variable stress. To the extent that they are successful in explaining variance, the power of the test for the effect of *ExpCon* will be increased. Therefore, although logically we can test if explanatory variables at the first level have random coefficients at the second or third level, and if explanatory variables at the second level have random coefficients at the third level, these possibilities are not pursued. We do test a model with a random coefficient for *ExpCon* at the third level, where there turns out to be significant slope variation. This varying slope can be predicted by adding a cross-level interaction between the variables *ExpCon* and *HospSize*. In view of this interaction, the variables *ExpCon* and *HospSize* have been centered on their overall mean. Table 2.5 presents the results for a series of models.

The equation for the first model, the intercept-only model, is:

$$stress_{ijk} = \gamma_{000} + v_{0k} + u_{0jk} + e_{ijk}. \tag{2.20}$$

This produces the variance estimates in the M0 column of Table 2.5. The proportion of variance (ICC) is .52 at the ward level, and .17 at the hospital level, calculated

*Table 2.5* Models for stress in hospitals and wards

| Model | M0: intercept only | M1: with predictors | M2: with random slope *ExpCon* | M3: with cross-level interaction |
|---|---|---|---|---|
| **Fixed part** | Coef. (s.e.) | Coef. (s.e.) | Coef. (s.e.) | Coef. (s.e.) |
| Intercept | 5.00 (0.11) | 5.50 (.12) | 5.46 (.12) | 5.50 (.11) |
| ExpCon | | –0.70 (.12) | –0.70 (.18) | –0.50 (.11) |
| Age | | 0.02 (.002) | 0.02 (.002) | 0.02 (.002) |
| Gender | | –0.45 (.04) | –0.46 (.04) | –0.46 (.04) |
| Experience | | –0.06 (.004) | –0.06 (.004) | –0.06 (.004) |
| Ward type | | 0.05 (.12) | 0.05 (.07) | 0.05 (.07) |
| HospSize | | 0.46 (.12) | 0.29 (.12) | –0.46 (.12) |
| ExpCon × HospSize | | | | 1.00 (.16) |
| **Random part** | | | | |
| $\sigma^2_{e\,ijk}$ | 0.30 (.01) | 0.22 (.01) | 0.22 (.01) | 0.22 (.01) |
| $\sigma^2_{u0jk}$ | 0.49 (.09) | 0.33 (.06) | 0.11 (.03) | 0.11 (.03) |
| $\sigma^2_{v0k}$ | 0.16 (.09) | 0.10 (.05) | 0.166 (.06) | 0.15 (.05) |
| $\sigma^2_{u1j}$ | | | 0.66 (.22) | 0.18 (.09) |
| **Deviance** | 1942.4 | 1604.4 | 1574.2 | 1550.8 |

following equations 2.16 and 2.17. The nurse-level and the ward-level variances are evidently significant. The test statistic for the hospital-level variance is $Z = 0.162/0.0852 = 1.901$, which produces a one-sided $p$-value of .029. The hospital-level variance is significant at the 5% level. The sequence of models in Table 2.5 shows that all predictor variables have a significant effect, except the ward type, and that the experimental intervention significantly lowers stress. The experimental effect varies across hospitals, and a large part of this variation can be explained by hospital size; in large hospitals the experimental effect is smaller.

## 2.5 A NOTE ABOUT NOTATION AND SOFTWARE

### 2.5.1 Notation

In general, there will be more than one explanatory variable at the lowest level and more than one explanatory variable at the highest level. Assume that we have *P* explanatory variables $X$ at the lowest level, indicated by the subscript $p$ ($p = 1 \ldots P$). Likewise, we have $Q$ explanatory variables $Z$ at the highest level, indicated by the subscript $q$ ($q = 1 \ldots Q$). Then, equation 2.5 becomes the more general equation:

$$Y_{ij} = \gamma_{00} + \gamma_{p0}X_{pij} + \gamma_{0q}\, Z_{qj} + \gamma_{pq}\, Z_{qj}X_{pij} + u_{pj}\, X_{pij} + u_{0j} + e_{ij}. \tag{2.21}$$

Using summation notation, we can express the same equation as:

$$Y_{ij} = \gamma_{00} + \sum_{p} \gamma_{p0}X_{pij} + \sum_{q} \gamma_{0q}Z_{qj} + \sum_{p}\sum_{q} \gamma_{pq}X_{pij}Z_{qj} + \sum_{p} u_{pj}X_{pij} + u_{0j} + e_{ij}. \tag{2.22}$$

The errors at the lowest level $e_{ij}$ are assumed to have a normal distribution with a mean of zero and a common variance $\sigma_e^2$ in all groups. The $u$-terms $u_{0j}$ and $u_{pj}$ are the residual error terms at the highest level. They are assumed to be independent from the errors $e_{ij}$ at the individual level, and to have a multivariate normal distribution with means of zero. The variance of the residual errors $u_{0j}$ is the variance of the intercepts between the groups, symbolized by $\sigma_{u_0}^2$. The variances of the residual errors $u_{pj}$ are the variances of the slopes between the groups, symbolized by $\sigma_{u_p}^2$. The *covariances* between the residual error terms $\sigma_{u_{pp}}$ are generally not assumed to be zero; they are collected in the higher-level variance/covariance matrix Ω.[8]

Note that in equation 2.22, $\gamma_{00}$, the regression coefficient for the intercept, is not associated with an explanatory variable. We can expand the equation by providing an explanatory variable that is a constant equal to one for all observed units. This yields the equation:

$$Y_{ij} = \gamma_{p0}\, X_{pij} + \gamma_{pq}\, Z_{qj}X_{pij} + u_{pj}\, X_{pij} + e_{ij} \tag{2.23}$$

where $X_{0ij} = 1$, and $p = 0 \ldots P$. Equation 2.23 makes clear that the intercept is a regression coefficient, just like the other regression coefficients in the equation. Some multilevel software, for instance HLM (Raudenbush, Bryk, Cheong, & Congdon, 2004), puts the intercept variable $X_0 = 1$ in the regression equation by default. Other multilevel software, for instance MLwiN (Rasbash, Steele, Browne, & Goldstein, 2009), requires that the analyst includes a variable in the data set that equals one in all cases, which must be added explicitly to the regression equation. In some cases, being able to eliminate the intercept term from the regression equation is a convenient feature.

[8] We may attach a subscript to Ω to indicate to which level it belongs. As long as there is no risk of confusion, the simpler notation without the subscript is used.

Equation 2.23 can be made very general if we let **X** be the matrix of all explanatory variables in the fixed part, symbolize the residual errors at all levels by $\mathbf{u}^{(l)}$ with $l$ denoting the level, and associate all error components with predictor variables **Z**, which may or may not be equal to the **X**. This produces the very general matrix formula $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}^{(l)}\mathbf{u}^{(l)}$ (see Goldstein, 1995, Appendix 2.1). Since this book is more about applications than about mathematical statistics, it generally uses the algebraic notation, except when multivariate procedures such as structural equation modeling are discussed.

The notation used in this book is close to the notation used by Goldstein (1987, 2003), Hox (1995), and Kreft and de Leeuw (1998). The most important difference is that these authors indicate the higher-level variance by $\sigma_{00}$ instead of our $\sigma^2_{u_0}$. The logic is that, if $\sigma_{01}$ indicates the covariance between variables 0 and 1, then $\sigma_{00}$ is the covariance of variable 0 with itself, which is its variance. Bryk and Raudenbush (1992) and Snijders and Bosker (1999) use a different notation; they denote the lowest-level error terms by $r_{ij}$, and the higher-level error terms by $u_j$. The lowest-level variance is $\sigma^2$ in their notation. The higher-level variances and covariances are indicated by the Greek letter *tau*; for instance, the intercept variance is given by $\tau_{00}$. The $\tau_{pp}$ are collected in the matrix Tau, symbolized as T. The HLM program and manual in part use a different notation, for instance when discussing longitudinal and three-level models.

In models with more than two levels, two different notational systems are used. One approach is to use different Greek characters for the regression coefficients at different levels, and different (Greek or Latin) characters for the variance terms at different levels. With many levels, this becomes cumbersome, and it is simpler to use the same character, say $\beta$ for the regression slopes and $u$ for the residual variance terms, and let the number of subscripts indicate to which level these belong.

### 2.5.2 Software

Multilevel models can be formulated in two ways: (1) by presenting separate equations for each of the levels, and (2) by combining all equations by substitution into a single model equation. The software HLM (Raudenbush et al., 2004) requires specification of the separate equations at each available level, but it can also show the single-equation version. Most other software, for example MLwiN (Rasbash et al., 2009), SAS Proc Mixed (Littell et al., 1996), SPSS Command Mixed (Norusis, 2005), uses the single-equation representation. Both representations have their advantages and disadvantages. The separate-equation representation has the advantage that it is always clear how the model is built up. The disadvantage is that it hides from view that modeling regression slopes by other variables results in adding an interaction to the model. As will be explained in Chapter 4, estimating and interpreting interactions correctly requires careful thinking. On the other hand, while the single-equation representation makes the existence of interactions obvious, it conceals the role of the complicated error components that are created by modeling varying slopes. In practice, to keep track of the model, it is recommended to start by writing the separate equations for the separate levels, and to use substitution to arrive at the single-equation representation.

To take a quote from Singer's excellent introduction to using SAS Proc Mixed for multilevel modeling (Singer, 1998, p. 350): 'Statistical software does not a statistician make. That said, without software, few statisticians and even fewer empirical researchers would fit the kinds of sophisticated models being promulgated today.' Indeed, software does not make a statistician, but the advent of powerful and user-friendly software for multilevel modeling has had a large impact in research fields as diverse as education, organizational research, demography, epidemiology, and medicine. This book focuses on the conceptual and statistical issues that arise in multilevel modeling of complex data structures. It assumes that researchers who apply these techniques have access to and familiarity with *some* software that can estimate these models. Software is mentioned in various places, especially when a technique is discussed that requires specific software features or is only available in a specific program.

Since statistical software evolves rapidly, with new versions of the software coming out much faster than new editions of general handbooks such as this, I do not discuss software setups or output in detail. As a result, this book is more about the possibilities offered by the various techniques than about how these things can be done in a specific software package. The techniques are explained using analyses on small but realistic data sets, with examples of how the results could be presented and discussed. At the same time, if the analysis requires that the software used have some specific capacities, these are pointed out. This should enable interested readers to determine whether their software meets these requirements, and assist them in working out the software setups for their favorite package.

In addition to the relevant program manuals, several software programs have been discussed in introductory articles. Using SAS Proc Mixed for multilevel and longitudinal data is discussed by Singer (1998). Peugh and Enders (2005) discuss SPSS Mixed using Singer's examples. Both Arnold (1992), and Heck and Thomas (2009) discuss multilevel modeling using HLM and Mplus as the software tool. Sullivan, Dukes, and Losina (1999) discuss HLM and SAS Proc Mixed. West, Welch, and Gatecki (2007) present a series of multilevel analyses using SAS, SPSS, R, Stata, and HLM. Finally, the multilevel modeling program at the University of Bristol maintains a multilevel homepage that contains a series of software reviews. The homepage for this book (on www.joophox.net) contains links to these and other multilevel resources.

The data used in the various examples are described in Appendix A, and are all available through the Internet.

# 3

# Estimation and Hypothesis Testing in Multilevel Regression

The usual method to estimate the values of the regression coefficients and the intercept and slope variances is the maximum likelihood method. This chapter gives a nontechnical explanation of this estimation method, to enable analysts to make informed decisions on the estimation options presented by present software. Some alternatives to maximum likelihood estimation are briefly discussed. Recent developments, such as bootstrapping and Bayesian estimation methods, are also briefly introduced in this chapter. In addition, these are explained in more detail in Chapter 13. Finally, this chapter describes some procedures that can be used to test hypotheses about specific parameters.

## 3.1 WHICH ESTIMATION METHOD?

Estimation of parameters (regression coefficients and variance components) in multilevel modeling is mostly done by the maximum likelihood method. The maximum likelihood (ML) method is a general estimation procedure, which produces estimates for the population parameters that maximize the probability (produce the 'maximum likelihood') of observing the data that are actually observed, given the model (see Eliason, 1993). Other estimation methods that have been used in multilevel modeling are generalized least squares (GLS), generalized estimating equations (GEE), and Bayesian methods such as Markov chain Monte Carlo (MCMC). Bootstrapping methods (see Mooney & Duval, 1993) can be used to improve the parameter estimates and the standard errors. In this section, I will discuss these methods briefly.

### 3.1.1 Maximum likelihood

Maximum likelihood (ML) is the most commonly used estimation method in multilevel modeling. An advantage of the maximum likelihood estimation method is that it is generally robust, and produces estimates that are asymptotically efficient and consistent. With large samples, ML estimates are usually robust against mild violations of the assumptions, such as having non-normal errors. Maximum likelihood estimation proceeds by maximizing a function called the likelihood function. Two different likelihood functions are used in multilevel regression modeling. One is full maximum likelihood (FML); in this method, both the regression coefficients and the variance components are included in the likelihood function. The other estimation method is restricted maximum likelihood (RML); here only the variance components are included in the likelihood function, and the regression coefficients are estimated in a second estimation step. Both methods produce parameter estimates with associated standard errors and an overall model *deviance*, which is a function of the likelihood. FML treats the regression coefficients as fixed but unknown quantities when the variance components are estimated, but does not take into account the degrees of freedom lost by estimating the fixed effects. RML estimates the variance components after removing the fixed effects from the model (see Searle, Casella, & McCulloch, 1992, Chapter 6). As a result, FML estimates of the variance components are biased; they are generally too small. RML estimates have less bias (Longford, 1993). RML also has the property that if the groups are balanced (have equal group sizes), the RML estimates are equivalent to analysis of variance (ANOVA) estimates, which are optimal (Searle et al., 1992, p. 254). Since RML is more realistic, it should, in theory, lead to better estimates, especially when the number of groups is small (Bryk & Raudenbush, 1992; Longford, 1993). In practice, the differences between the two methods are usually small (see Hox, 1998; Kreft & de Leeuw, 1998). For example, if we compare the FML estimates for the intercept-only model for the popularity data in Table 2.1 with the corresponding RML estimates, the only difference within two decimals is the intercept variance at level 2. FML estimates this as 0.69, and RML as 0.70. The size of this difference is absolutely trivial. If nontrivial differences are found, the RML method usually performs better (Browne, 1998). FML still continues to be used, because it has two advantages over RML. First, the computations are generally easier, and second, since the regression coefficients are included in the likelihood function, an overall chi-square test based on the likelihood can be used to compare two models that differ in the fixed part (the regression coefficients). With RML, only differences in the random part (the variance components) can be compared with this test. Most tables in this book have been produced using FML estimation, if RML is used this is explicitly stated in the text.

Computing the maximum likelihood estimates requires an *iterative* procedure. At the start, the computer program generates reasonable starting values for the various parameters (in multilevel regression analysis these are usually based on single-level regression estimates). In the next step, an ingenious computation procedure tries to improve on the starting values, to produce better estimates. This second step is repeated (iterated) many times. After each iteration, the program inspects how much the estimates have actually changed compared to the previous step. If the changes are very small, the program concludes that the estimation procedure has *converged* and that it is finished. Using multilevel software, we generally take the computational details for granted. However, computational problems do sometimes occur. A problem common to programs using an iterative maximum likelihood procedure is that the iterative process is not always *guaranteed* to stop. There are models and data sets for which the program may go through an endless sequence of iterations, which can only be ended by stopping the program. Because of this, most programs set a built-in limit for the maximum number of iterations. If convergence is not reached within this limit, the computations can be repeated with a higher limit. If the computations do not converge after an extremely large number of iterations, we suspect that they may never converge.[1] The problem is how one should interpret a model that does not converge. The usual interpretation is that a model for which convergence cannot be reached is a bad model, using the simple argument that if estimates cannot be found, this disqualifies the model. However, the problem may also lie with the data. Especially with small samples, the estimation procedure may fail even if the model is valid. In addition, it is even possible that, if only we had a better computer algorithm, or better starting values, we could find acceptable estimates. Still, experience shows that if a program does not converge with a data set of reasonable size, the problem often is a badly misspecified model. In multilevel analysis, non-convergence often occurs when we try to estimate too many random (variance) components that are actually close or equal to zero. The solution is to simplify the model by leaving out some random components; often the estimated values from the non-converged solution provide an indication of which random components can be omitted.

### 3.1.2 Generalized least squares

Generalized least squares (GLS) is an extension of the standard estimation ordinary least squares (OLS) method that allows for heterogeneity and observations that differ in sampling variance. GLS estimates approximate ML estimates, and they are asymptotically equivalent. Asymptotic equivalence means that in very large samples they are in practice indistinguishable. Goldstein (2003, p. 21) notes that 'expected GLS' estimates can be obtained from a maximum likelihood procedure by restricting the number of iterations to one. Since GLS estimates are obviously faster to compute than full ML estimates, they can be used as a stand-in for ML estimates in computationally intensive procedures such as extremely large data sets, or when bootstrapping is used. They can also be used when ML procedures fail to converge; inspecting the GLS results may help to diagnose the problem. Furthermore, since GLS estimates are respectable statistical estimates in their own right, in such situations one can report the GLS estimates instead of the more usual ML estimates. However, simulation research

[1] Some programs allow the analyst to monitor the iterations, to observe whether the computations are going somewhere, or are just moving back and forth without improving the likelihood function.

shows that, in general, GLS estimates are less efficient, and the GLS-derived standard errors are rather inaccurate (see Hox, 1998; Kreft, 1996; van der Leeden, Meijer, & Busing, 2008). Therefore, in general, ML estimation should be preferred.

### 3.1.3 Generalized estimating equations

The generalized estimating equations method (GEE, see Liang & Zeger, 1986) estimates the variances and covariances in the random part of the multilevel model directly from the residuals, which makes them faster to compute than full ML estimates. Typically, the dependences in the multilevel data are accounted for by a very simple model, represented by a *working correlation matrix*. For individuals within groups, the simplest assumption is that the respondents within the same group all have the same correlation. For repeated measures, a simple autocorrelation structure is usually assumed. After the estimates for the variance components are obtained, GLS is used to estimate the fixed regression coefficients. Robust standard errors are generally used to counteract the approximate estimation of the random structure. For non-normal data this results in a *population-average model*, where the emphasis is on estimating average population effects and not on modeling individual and group differences. Raudenbush and Bryk (2002) describe the multilevel *unit-specific* model (usually based on ML estimation) as a model that aims to model the effect of predictor variables while controlling statistically for other predictor variables at different levels, plus the random effects in the model. In contrast, the *population-average* model (usually based on GEE estimation) controls for the other predictor variables, but not for the random effects. When a nonlinear model is estimated, the GEE estimates are different from the ML estimates. For example, in an intercept-only logistic regression model the average probability in the population of repeating a class can be calculated from the population-average estimate of the intercept. The unit-specific intercept can in general not be used to calculate this probability. If the interest is group-level variation, for instance in modeling the differences in the level 1 effects using level 2 variables, the unit-specific model is appropriate. If we are only interested in population estimates of the average effect of level 1 variables, for instance in the difference between boys and girls nationwide in the probability of repeating a class, the population-average model is appropriate. For a further discussion, I refer to Zeger, Liang, and Albert (1988) and Hu, Goldberg, Hedeker, Flay, and Pentz (1998).

According to Goldstein (2003) and Raudenbush and Bryk (2002), GEE estimates are less efficient than full ML estimates, but they make weaker assumptions about the structure of the random part of the multilevel model. If the model for the random part is correctly specified, ML estimators are more efficient, and the model-based (ML) standard errors are generally smaller than the GEE-based robust standard errors. If the model for the random part is incorrect, the GEE-based estimates and robust standard errors are still consistent. So, provided the sample size is reasonably large, GEE estimators are robust against misspecification of the random part of the model, including violations of the normality assumption. A drawback of the GEE approach is that it only approximates the random effects structure, and therefore the random effects cannot be analyzed in detail. So, most software will estimate a full unstructured covariance matrix for the random part, which makes it impossible to estimate random effects for the intercept or slopes. Given the general robustness of ML methods, it is preferable to use ML methods when these are available, and to use robust estimators or bootstrap corrections when there is serious doubt about the assumptions of the ML method. Robust estimators, which are related to GEE estimators (Burton, Gurrin, & Sly, 1998), and bootstrapping are treated in more detail in Chapter 13 of this book.

### 3.1.4 Bootstrapping

In bootstrapping, random samples are repeatedly drawn with replacement from the observed data. In each of these random samples, the model parameters are estimated, generally using either FML or RML maximum likelihood estimation. This process is repeated *b* times. For each model parameter, this results in a set of *b* parameter estimates. The variance of these *b* estimates is used as an indicator of the sampling variance associated with the parameter estimate obtained from the full sample. Since the bootstrap samples are obtained by resampling from the total sample, bootstrapping falls under the general term of resampling methods (see Good, 1999). Bootstrapping can be used to improve both the point estimates and the standard errors. Typically, at least 1000 bootstrap samples are needed for sufficient accuracy. This makes the method computationally demanding, but less so than the Bayesian methods treated in the next section. Since bootstrapping has its own complications, it is discussed in more detail in Chapter 13. If we execute a bootstrap estimation for our example data, the results are almost identical to the asymptotic FML results reported in Table 2.2. The estimates differ by 0.01 at most, which is a completely trivial difference. Bootstrap estimates are most attractive when we have reasons to suspect the asymptotic results, for example because we have a small sample size, or because we have non-normal data.

### 3.1.5 Bayesian methods

In Bayesian statistics, we express our uncertainty about the population values of the model parameters by assigning to them a distribution of possible values. This distribution is called the *prior* distribution, because it is specified independently from the data. The prior distribution is combined with the likelihood of the data to produce a *posterior* distribution, which describes our uncertainty about the population values after observing the data. Typically, the variance of the posterior distribution is smaller than the variance of the prior distribution, which means that observing the data has reduced our uncertainty about the possible population values. For the prior distribution, we have a fundamental choice between using an informative prior and using an uninformative prior. An informative prior is a peaked distribution with a small variance, which expresses a strong belief about the unknown population parameter. An informative prior will, of course, strongly influence the posterior distribution, and hence our conclusions. For this reason, many statisticians prefer an uninformative or diffuse prior, which has very little influence on the posterior, and only serves to produce the posterior. An example of an uninformative prior is the uniform distribution, which simply states that the unknown parameter value is a number between minus and plus infinity, with all values equally likely.

If the posterior distribution has a mathematically simple form, for instance a normal distribution, we can use this distribution to produce a point estimate and a confidence interval for the population parameter. However, in complex multivariate models, the posterior is generally a complicated multivariate distribution, which makes it difficult to use it directly to produce parameter estimates and confidence intervals. Therefore, simulation techniques are used to generate random samples from the posterior distribution. The simulated posterior distribution is then used to provide a point estimate (typically the mode or median of the simulated values) and a confidence interval.

Bayesian methods can provide accurate estimates of the parameters and the uncertainty associated with them (Goldstein, 2003). However, they are computationally demanding, and the simulation procedure must be monitored to insure that it is working properly. Bayesian estimation methods are treated in more detail in Chapter 13.

## 3.2 SIGNIFICANCE TESTING AND CONFIDENCE INTERVALS

This section discusses procedures for testing significance and constructing confidence intervals for the regression coefficients and variance components. There is also a global test based on the likelihood that can be used to compare two nested models. Finally, contrasts of regression coefficients or variances are described, and constraints that can be used to constrain regression coefficients or variance components to specific values.

### 3.2.1 Testing regression coefficients and variance components

Maximum likelihood estimation produces parameter estimates and corresponding standard errors. These can be used to carry out a significance test of the form $Z =$

(*estimate*)/(*standard error of estimate*), where $Z$ is referred to the standard normal distribution. This test is known as the *Wald test* (Wald, 1943). The standard errors are asymptotic, which means that they are valid for large samples. As usual, it is not precisely known when a sample is large enough to give confidence about the precision of the estimates. Simulation research suggests that for accurate standard errors for level 2 variances, a relatively large level 2 sample size is needed. For instance, simulations by van der Leeden, Busing, and Meijer (1997) suggest that with fewer than 100 groups, ML estimates of variances and their standard errors are not very accurate. In ordinary regression analysis, a rule of thumb is to require $104 + p$ observations if the interest is in estimating and interpreting regression coefficients, where $p$ is the number of explanatory variables (Green, 1991). If the interest is in interpreting (explained) variance, the rule of thumb is to require $50 + 8p$ observations. In multilevel regression, the relevant sample size for higher-level coefficients and variance components is the number of groups, which is often not very large. Green's rule of thumb and van der Leeden et al.'s simulation results agree on a preferred group-level sample size of at least 100. Additional simulation research (Maas & Hox, 2005) suggests that if the interest lies mostly in the fixed part of the model, far fewer groups are sufficient, especially for the lowest-level regression coefficients. The issue of the sample sizes needed to produce accurate estimates and standard errors is taken up in more detail in Chapter 12.

It should be noted that the $p$-values and confidence intervals produced by the program HLM (Raudenbush, Yang, & Yosef, 2000) differ from those obtained from most other programs. As part of their output, most multilevel analysis programs produce parameter estimates and asymptotic standard errors for these estimates, all obtained from the maximum likelihood estimation procedure. The usual significance test is the Wald test, with $Z$ evaluated against the standard normal distribution. Bryk and Raudenbush (1992, p. 50), referring to a simulation study by Fotiu (1989), argue that for the fixed effects it is better to refer this ratio to a $t$-distribution on $J - p - 1$ degrees of freedom, where $J$ is the number of second level units, and $p$ is the total number of explanatory variables in the model. The $p$-values produced by the program HLM (Raudenbush, Bryk, Cheong, & Congdon, 2004) are based on these tests rather than on the more common Wald tests. When the number of groups $J$ is large, the difference between the asymptotic Wald test and the alternative Student $t$-test is very small. However, when the number of groups is small, the differences may become important. Since referring the result of the $Z$-test on the regression coefficients to a Student $t$-distribution is conservative, this procedure should provide a better protection against type I errors. A better choice for the degrees of freedom in multilevel models is provided by the Satterthwaite approximation (Sattertwaithe, 1946). This estimates the number of degrees of freedom using the values of the residual variances. Simulation research (Manor & Zucker, 2004) shows that the Satterthwaite approximation works better than the Wald test when the sample size is small (e.g., smaller than 30).

Several authors (e.g., Berkhof & Snijders, 2001; Raudenbush & Bryk, 2002) argue that the *Z*-test is not appropriate for the variances, because it assumes a normal distribution, whereas the sampling distribution of variances is skewed, especially when the variance is small. Especially if we have both a small sample of groups and a variance component close to zero, the distribution of the Wald statistic is clearly non-normal. Raudenbush and Bryk propose to test variance components using a chi-square test on the residuals. This chi-square is computed by:

$$\chi^2 = \sum(\hat{\beta}_j - \beta)^2 \;/\; \hat{V}_j, \tag{3.1}$$

where $\hat{\beta}_j$ is the OLS estimate of a regression coefficient computed separately in group *j*, $\beta$ its overall estimate, and $\hat{V}_j$ its estimated sampling variance in group *j*. The number of degrees of freedom is given by $df = J - p - 1$, where *J* is the number of second-level units, and *p* is the total number of explanatory variables in the model. Groups that have a small number of cases are passed over in this test, because their OLS estimates are not sufficiently accurate.

Simulation studies on the Wald test for a variance component (van der Leeden et al., 1997) and the alternative chi-square test (Harwell, 1997; Sánchez-Meca & Marín-Martínez, 1997) suggest that with small numbers of groups, both tests suffer from a very low power. The test that compares a model with and without the parameters under consideration, using the chi-square model test described in the next section, is generally better (Berkhof & Snijders, 2001; Goldstein, 2003). Only if the likelihood is determined with a low precision, which is the case for some approaches to modeling non-normal data, is the Wald test preferred. Note that if the Wald test is used to test a variance component, a one-sided test is the appropriate one.

### 3.2.2 Comparing nested models

From the likelihood function we can calculate a statistic called the *deviance* that indicates how well the model fits the data. The deviance is defined as −2 × ln (*Likelihood*), where *Likelihood* is the value of the likelihood function at convergence, and is the natural logarithm. In general, models with a lower deviance fit better than models with a higher deviance. If two models are *nested*, which means that a specific model can be derived from a more general model by removing parameters from the general model, we can compare them statistically using their deviances. The difference of the deviances for two nested models has a chi-square distribution, with degrees of freedom equal to the difference in the number of parameters estimated in the two models. This can be used to perform a formal chi-square test to test whether the more general model fits significantly better than the simpler model. The deviance difference test is also referred to as the likelihood ratio test, since the ratio of two likelihoods is compared by looking at the difference of their logarithms.

The chi-square test of the deviances can be used to good effect to explore the importance of random effects, by comparing a model that contains these effects with a model that excludes them.

Table 3.1 presents two models for the pupil popularity data used as an example in Chapter 2. The first model contains only an intercept. The second model adds two pupil-level variables and a teacher-level variable, with the pupil-level variable extraversion having random slopes at the second (class) level. To test the second-level variance component $\sigma_{u0}^2$ using the deviance difference test, we remove it from model M0. The resulting model (not presented in Table 3.1) produces a deviance of 6970.4, and the deviance difference is 642.9. Since the modified model estimates one parameter less, this is referred to the chi-square distribution with one degree of freedom. The result is obviously significant.

*Table 3.1* Intercept-only model and model with explanatory variables

| Model | M0: intercept only | M1: with predictors |
|---|---|---|
| **Fixed part** | Coefficient (s.e.) | Coefficient (s.e.) |
| Intercept | 5.08 (.09) | 0.74 (.20) |
| Pupil gender | | 1.25 (.04) |
| Pupil extraversion | | 0.45 (.02) |
| Teacher experience | | 0.09 (.01) |
| **Random part** | | |
| $\sigma_e^2$ | 1.22 (.04) | 0.55 (.02) |
| $\sigma_{u0}^2$ | 0.69 (.11) | 1.28 (.28) |
| $\sigma_{u02}$ | | –0.18 (.05) |
| $\sigma_{u2}^2$ | | 0.03 (.008) |
| **Deviance** | 6327.5 | 4812.8 |

The variance of the regression coefficient for pupil gender is estimated as zero, and therefore it is removed from the model. A formal test is not necessary. In model M1 in Table 3.1 this variable is treated as fixed: no variance component is estimated. To test the significance of the variance of the extraversion slopes, we must remove the variance parameter from the model. This presents us with a problem, since there is also a covariance parameter $\sigma_{u02}$ associated with the extraversion slopes. If we remove both the variance and the covariance parameter from the model, we are testing a combined hypothesis on two degrees of freedom. It is better to separate these tests. Some software (e.g., MLwiN) actually allows us to remove the variance of the slopes from the model but to retain the covariance parameter. This is a strange model, but for testing purposes it allows us to carry out a separate test on the variance parameter only. Other software (e.g., MLwiN, SPSS, SAS) allows the removal of the covariance parameter, while keeping the variance in the model. If we modify model M1 this way, the deviance increases to 4851.9. The difference is 39.1, which is a chi-square variate with one degree of freedom, and highly significant. If we modify the model further, by removing the slope variance, the deviance increases again to 4862.3. The difference with the previous model is 10.4, again with one degree of freedom, and it is highly significant.

Asymptotically, the Wald test and the test using the chi-square difference are equivalent. In practice, the Wald test and the chi-square difference test do not always lead to the same conclusion. If a variance component is tested, the chi-square difference test is clearly better, except when models are estimated where the likelihood function is only an approximation, as in the logistic models discussed in Chapter 6.

When the chi-square difference test is used to test a variance component, it should be noted that the standard application leads to a *p*-value that is too high. The reason is that the null-hypothesis of zero variance is on the boundary of the parameter space (all possible parameter values) since variances cannot be negative. If the null-hypothesis is true, there is a 50% chance of finding a positive variance, and a 50% chance of finding a negative variance. Negative variances are inadmissible, and the usual procedure is to change the negative estimate to zero. Thus, under the null-hypothesis the chi-square statistic has a mixture distribution of 50% zero and 50% chi-square with one degree of freedom. Therefore, the *p*-value from the chi-square difference test must be divided by two if a variance component is tested (Berkhof & Snijders, 2001). If we test a slope variance, and remove both the slope variance and the covariance from the model, the mixture is more complicated, because we have a mixture of 50% chi-square with one degree of freedom for the unconstrained intercept–slope covariance and 50% chi-square with two degrees of freedom for the covariance and the variance that is constrained to be non-negative (Verbeke & Molenberghs, 2000). The *p*-value for this mixture is calculated using $p = 0.5\,P\left(\chi_1^2 > C^2\right) + 0.5P\left(\chi_2^2 > C^2\right)$ where $C^2$ is the difference in the deviances of the model with and without the slope variance and intercept–slope covariance. Stoel, Galindo, Dolan, and van den Wittenboer (2006) discuss how to carry out such tests in general. If it is possible to remove the intercept–slope covariance from the model, it is possible to test the significance of the slope variance with a one degree of freedom test, and we can simply halve the *p*-value again. For the regression coefficients, the chi-square test (only in combination with FML estimation) is in general also superior. The reason is that the Wald test is to some degree sensitive to the parameterization of the model and the specific restrictions to be tested (Davidson & MacKinnon, 1993, Chapter 13.5–13.6). The chi-square test is invariant under reparameterizations of the model. Since the Wald test is much more convenient, it is in practice used the most, especially for the fixed effects. Even so, if there is a discrepancy between the result of a chi-square difference test and the equivalent Wald test, the chi-square difference test is generally the preferred one.

LaHuis and Ferguson (2009) compare, among others, the chi-square deviance test and the chi-square residuals test described above. In their simulation, all tests controlled the type I error well, and the deviance difference test (dividing *p* by two, as described above) generally performed best in terms of power.

### 3.2.3 Comparing non-nested models

If the models to be compared are not nested models, the principle that models should be as simple as possible (theories and models should be parsimonious) indicates that we should generally keep the simpler model. A general fit index to compare the fit of statistical models is Akaike's Information Criterion, *AIC* (Akaike, 1987), which was developed to compare non-nested models, adjusting for the number of parameters estimated. The AIC for multilevel regression models is conveniently calculated from the deviance *d*, and the number of estimated parameters *q*:

$$AIC = d + 2q. \tag{3.2}$$

The AIC is a very general fit index that assumes that we are comparing models that are fit to the same data set, using the same estimation method. A fit index similar to the AIC is Schwarz's Bayesian Information Criterion, *BIC* (Schwarz, 1978), which is given by:

$$BIC = d + q\ \ln(N). \tag{3.3}$$

In multilevel modeling, the general equation 3.3 for the BIC is ambiguous, because it is unclear whether *N* refers to the first-level or the second-level sample size. What *N* means in equation 3.3 is differently chosen by different software. Most software uses the number of units at the highest level for the *N*. This makes sense when multilevel models are used for longitudinal data, where the highest level is often the subject level. Given the strong interest in multilevel modeling in contextual effects, choosing the highest-level sample size appears a sensible rule.

When the deviance goes down, indicating a better fit, both the AIC and the BIC also tend to go down. However, the AIC and the BIC both include a penalty function based on the number of estimated parameters *q*. As a result, when the number of estimated parameters goes up, the AIC and BIC tend to go up too. For most sample sizes, the BIC places a larger penalty on complex models, which leads to a preference for smaller models. Since multilevel data have a different sample size at different levels, the AIC is more straightforward than the BIC, and is therefore the recommended choice. The AIC and BIC are typically used to compare a range of competing models, and the model(s) with the lowest AIC or BIC value is considered the most attractive. Both the AIC and the BIC have been shown to perform well, with a small advantage for the BIC (Haughton, Oud, & Jansen, 1997; Kieseppä, 2003). It should be noted that the AIC and BIC are based on the likelihood function. With FML estimation, the AIC and BIC can be used to compare models that differ either in the fixed part or in the random part. If RML estimation is used, it can only be used to compare models that differ in the random part. Since RML effectively partials out the fixed part, before the random part is estimated, the RML likelihood may still change if the fixed part is changed. Therefore, if likelihood-based procedures are used to compare models using RML estimation, the fixed part of the model should be kept constant. Most current software does not produce the AIC or BIC (SPSS and SAS do), but they can be calculated using the formulas given earlier. For an introductory discussion of the background of the AIC and the BIC see McCoach and Black (2008).

## 3.3 CONTRASTS AND CONSTRAINTS

It is possible to define a *contrast* for a set of regression coefficients or variance components. A contrast is a composite hypothesis on a set of parameters, for example, that they are equal, or that a set of parameters are all equal to zero. In general, a contrast has the form:

$$H_0\text{: } \mathbf{Cp} = \mathbf{k}, \tag{3.4}$$

where **C** is a contrast matrix that specifies the composite hypothesis, **p** is the vector of parameters that are involved, and **k** is a vector of contrast values. An example makes clear how this works. Table 3.1 presents the results of analyzing the popularity data. There were 2000 pupils in 100 classes. The intercept-only model lists two random parameters: the intercept variance at level 1, which is estimated as 1.22, and the intercept variance at level 2, which is estimated as 0.69. Suppose that we want to evaluate the null-hypothesis that these variance components are equal. Thus, the null-hypothesis is: $\sigma_e^2 - \sigma_u^2 = 0$. The parameter vector **p** is:

$$\mathbf{p} = \begin{Bmatrix} \sigma_e^2 \\ \sigma_{u0}^2 \end{Bmatrix}. \tag{3.5}$$

We can write the composite null-hypothesis as:

$$\{1, -1\}\begin{Bmatrix}\sigma_e^2 \\ \sigma_{u0}^2\end{Bmatrix} = \{0\}, \tag{3.6}$$

or, in simple algebraic notation:

$$1 \times \sigma_e^2 - 1 \times \sigma_{u0}^2 = 0. \tag{3.7}$$

Contrasts are tested using an asymptotic Wald chi-square test (Goldstein, 2003; Raudenbush & Bryk, 2002). If we test the null-hypothesis represented by contrast 3.6 on the example data, we obtain a chi-square of 21.1 with one degree of freedom, and a $p$-value of < .001. The null-hypothesis of equal variances at both levels is rejected.

Sometimes we have a composite hypothesis that is more complicated, and we may require more than one contrast vector. For example, suppose that we want to test the null-hypothesis that all three of the regression coefficients in Table 3.1 are equal. The list of all fixed coefficients is: the intercept, the regression coefficients for pupil extraversion and gender, and the regression coefficient for teacher experience. Our composite null-hypothesis is $H_0$: $\gamma_1 = \gamma_2 = \gamma_3$. This composite null-hypothesis can be expressed by having two simple null-hypotheses: $\gamma_1 - \gamma_2 = 0$ and $\gamma_2 - \gamma_3 = 0$. In matrix formulation, replacing the gammas by the corresponding variable names, we have:

$$\begin{Bmatrix}0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1\end{Bmatrix}\begin{Bmatrix}\textit{intercept} \\ \textit{pup. extrav.} \\ \textit{pup. gender} \\ \textit{teach. exper.}\end{Bmatrix} = \begin{Bmatrix}0 \\ 0\end{Bmatrix}. \tag{3.8}$$

Equation 3.8 represents the two contrasts given by $0 \times \gamma_{00} + 1 \times \gamma_{01} + (-1) \times \gamma_{02} + 0 \times \gamma_{10} = 0$ and $0 \times \gamma_{00} + 0 \times \gamma_{01} + 1 \times \gamma_{02} + (-1) \times \gamma_{10} = 0$, or more simply $\gamma_{01} = \gamma_{02}$ and $\gamma_{02} = \gamma_{10}$. If we test this composite null-hypothesis on the example data, we obtain a chi-square of 313.8 with two degrees of freedom (one for each row of the contrast matrix), and a $p$-value of < .001. The null hypothesis is clearly untenable.

In these two examples, the null-hypothesis is not very interesting. Contrast tests on parameters are theoretically informative when we can use them to compare regression coefficients or variance components that refer to similar variables. For instance, in a multivariate multilevel analysis, where we have two or more outcome variables, we can use contrasts to test if a predictor variable has the same effect on both outcome variables. Alternatively, if we have two comparable predictor variables, we can test if they both have the same effect on a single outcome variable. Contrast tests are also useful when categorical variables are represented by a set of dummies. Using complex contrasts, we can replace a set of separate Wald tests by one single omnibus test on the full set of dummies.

Some software allows putting *constraints* on a set of parameters. Simple constraints are constraining a single parameter to be equal to zero or one. This is used, for instance, in multilevel models for meta-analysis, which is treated in Chapter 11. Complex constraints are similar to complex contrasts, and allow constraining a set of parameters to be equal. A useful application is in models with dummy variables that identify groups. We can use a constraint matrix to specify that the variances for the group identification must be equal. This is used in models for cross-classified data, which is treated in Chapter 9.

Constraints are similar to contrasts, but do not specify a value for the contrast. If we compare models with and without constraints using the deviance chi-square test, we have a test that is asymptotically identical to the equivalent composite test using contrasts. The composite contrast test on model parameters is a composite Wald chi-square test (see Goldstein, 2003). The global model comparison using deviances is a likelihood ratio test. With normal data, the global likelihood ratio test is generally considered more accurate, especially when variance components are tested. With non-normal data, the likelihood is often approximated, and then the composite Wald test is preferred.

# 4

# Some Important Methodological and Statistical Issues

The multilevel regression model is more complicated than the standard single-level multiple regression model. One difference is the number of parameters, which is much larger in the multilevel model. This poses problems when models are fitted that have many parameters, and also in model exploration. Another difference is that multilevel models often contain interaction effects in the form of cross-level interactions. Interaction effects are tricky, and analysts should deal with them carefully. Finally, the multilevel model contains several different residual variances, and no single number can be interpreted as *the* amount of explained variance. These issues are treated in this chapter.

## 4.1 ANALYSIS STRATEGY

The number of parameters in a multilevel regression model can easily become very large. If there are *p* explanatory variables at the lowest level and *q* explanatory variables at the highest level, the multilevel regression model for two levels is given by equation 4.1:

$$Y_{ij} = \gamma_{00} + \gamma_{p0}\, X_{pij} + \gamma_{0q}\, Z_{qj} + \gamma_{pq}\, Z_{qj}X_{pij} + u_{pj}\, X_{pij} + u_{0j} + e_{ij}. \tag{4.1}$$

The number of estimated parameters in the model described by equation 4.1 is given by the following list:

| *Parameters* | *Number* |
|---|---|
| Intercept | 1 |
| Lowest-level error variance | 1 |
| Fixed slopes for the lowest-level predictors | $p$ |
| Highest-level error variance | 1 |
| Highest-level error variances for these slopes | $p$ |
| Highest-level covariances of the intercept with all slopes | $p$ |
| Highest-level covariances between all slopes | $p(p - 1)/2$ |
| Fixed slopes for the highest-level predictors | $q$ |
| Fixed slopes for cross-level interactions | $p \times q$ |

An ordinary single-level regression model for the same data would estimate only the intercept, one error variance, and $p + q$ regression slopes. The superiority of the multilevel regression model is clear, if we consider that the data are clustered in groups. If we have 100 groups, estimating an ordinary multiple regression model in each group separately requires estimating 100 × (1 regression intercept + 1 residual variance + $p$ regression slopes) plus possible interactions with the $q$ group-level variables. Multilevel regression replaces estimating 100 intercepts by estimating an average intercept plus its residual variance across groups, assuming a normal distribution for these residuals. Thus, multilevel regression analysis replaces estimating 100 separate intercepts by estimating two parameters (the mean and variance of the intercepts), plus a normality assumption. The same simplification is used for the regression slopes. Instead of estimating 100 slopes for the explanatory variable pupil gender, we estimate the average slope along with its variance across groups, and assume that the distribution of the slopes is normal. Nevertheless, even with a modest number of explanatory variables, multilevel regression analysis implies a complicated model. Generally, we do not want to estimate the complete model, first because this is likely to get us into computational problems, but also because it is very difficult to interpret such a complex model. We prefer more limited models that include only those parameters that have proven their worth in previous research, or are of special interest for our theoretical problem.

If we have no strong theories, we can use an exploratory procedure to select a model. Model building strategies can be either top-down or bottom-up. The top-down approach starts with a model that includes the maximum number of fixed and random effects that are considered for the model. Typically, this is done in two steps. The first step starts with all the fixed effects and possible interactions in the model, followed by removing insignificant effects. The second step starts with a rich random structure, followed by removal of insignificant effects. This procedure is described by West et al.

(2007). In multilevel modeling, the top-down approach has the disadvantage that it starts with a large and complicated model, which leads to longer computation time and sometimes to convergence problems. In this book, the opposite strategy is mostly used, which is bottom-up: start with a simple model and proceed by adding parameters, which are tested for significance after they have been added. Typically, the procedure starts by building up the fixed part, and follows after with the random part. The advantage of the bottom-up procedure is that it tends to keep the models simple.

It is attractive to start with the simplest possible model, the intercept-only model, and to add the various types of parameters step by step. At each step, we inspect the estimates and standard errors to see which parameters are significant, and how much residual error is left at the distinct levels. Since we have larger sample sizes at the lowest level, it makes sense to build up the model from there. In addition, since fixed parameters are typically estimated with much more precision than random parameters, we start with the fixed regression coefficients, and add variance components at a later stage. The different steps of such a selection procedure are given below.

*Step 1:*

Analyze a model with no explanatory variables. This model, the *intercept-only model*, is given by the model of equation 2.8, which is repeated here:

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij}. \tag{4.2}$$

In equation 4.2, $\gamma_{00}$ is the regression intercept, and $u_{0j}$ and $e_{ij}$ are the usual residuals at the group and the individual level. The intercept-only model is useful because it gives us an estimate of the intraclass correlation $\rho$:

$$\sigma^2_{u0} \Big/ \left(\sigma^2_{u0} + \sigma^2_e\right), \tag{4.3}$$

where $\sigma^2_{u0}$ is the variance of the group-level residuals $u_{0j}$, and $\sigma^2_e$ is the variance of the individual-level residuals $e_{ij}$. The intercept-only model also gives us a benchmark value of the deviance, which is a measure of the degree of misfit of the model, and which can be used to compare models as described in Chapter 3.

*Step 2:*

Analyze a model with all lower-level explanatory variables fixed. This means that the corresponding variance components of the slopes are fixed at zero. This model is written as:

$$Y_{ij} = \gamma_{00} + \gamma_{p0}\, X_{pij} + u_{0j} + e_{ij}, \tag{4.4}$$

where the $X_{pij}$ are the $p$ explanatory variables at the individual level. In this step, we assess the contribution of each individual-level explanatory variable. The significance of each predictor can be tested, and we can assess what changes occur in the first-level and second-level variance terms. If we use the FML estimation method, we can test the improvement of the final model chosen in this step by computing the difference of the deviance of this model and the previous model (the intercept-only model). This difference approximates a chi-square with, as degrees of freedom, the difference in the number of parameters of both models (see 3.1.1). In this case, the degrees of freedom are simply the number of explanatory variables added in step 2.

*Step 3:*
Add the higher-level explanatory variables:

$$Y_{ij} = \gamma_{00} + \gamma_{p0}\,X_{pij} + \gamma_{0q}\,Z_{qj} + u_{0j} + e_{ij} \tag{4.5}$$

where the $Z_{qj}$ are the $q$ explanatory variables at the group level. This model allows us to examine whether the group-level explanatory variables explain between-group variation in the dependent variable. Again, if we use FML estimation, we can use the global chi-square test to formally test the improvement of fit. If there are more than two levels, this step is repeated on a level-by-level basis.

The models in steps 2 and 3 are often denoted as *variance component* models, because they decompose the intercept variance into different variance components for each hierarchical level. In a variance component model, the regression intercept is assumed to vary across the groups, but the regression slopes are assumed fixed. If there are no higher-level explanatory variables, this model is equivalent to a random effects analysis of covariance (ANCOVA); the grouping variable is the usual ANCOVA factor, and the lowest-level explanatory variables are the covariates (see Kreft & de Leeuw, 1998, p. 30; Raudenbush & Bryk, 2002, p. 25). There is a difference in estimation method: ANCOVA uses OLS techniques and multilevel regression uses ML estimation. Nevertheless, both models are highly similar, and if the groups have all equal sizes, the model is equivalent to analysis of covariance. It is even possible to compute the usual ANCOVA statistics from the multilevel program output (Raudenbush, 1993a). The reason to start with models that include only fixed regression coefficients is that we generally have more information on these coefficients; they can be estimated with more precision than the variance components. When we are confident that we have a well-fitting model for the fixed part, we turn to modeling the random part.

*Step 4:*
Assess whether any of the slopes of any of the explanatory variables has a significant variance component between the groups. This model, the *random coefficient model*, is given by:

$$Y_{ij} = \gamma_{00} + \gamma_{p0}\, X_{pij} + \gamma_{0q}\, Z_{qj} + u_{pj}\, X_{pij} + u_{0j} + e_{ij} \tag{4.6}$$

where the $u_{pj}$ are the group-level residuals of the slopes of the individual-level explanatory variables $X_{pij}$.

Testing for random slope variation is best done on a variable-by-variable basis. When we start by including all possible variance terms in a model (which involves also adding many covariance terms), the result is most likely an overparameterized model with serious estimation problems, such as convergence problems or extremely slow computations. Variables that were omitted in step 2 may be analyzed again in this step; it is quite possible for an explanatory variable to have no significant average regression slope (as tested in step 2), but to have a significant variance component for this slope.

After deciding which of the slopes have a significant variance between groups, preferably using the deviance difference test, we add all these variance components simultaneously in a final model, and use the chi-square test based on the deviances to test whether the final model of step 4 fits better than the final model of step 3. Since we are now introducing changes in the random part of the model, the chi-square test can also be used with RML estimation (see 3.1.1). When counting the number of parameters added, remember that adding slope variances in step 4 also adds the covariances between the slopes!

If there are more than two levels, this step is repeated on a level-by-level basis.

*Step 5:*

Add cross-level interactions between explanatory group-level variables and those individual-level explanatory variables that had significant slope variation in step 4. This leads to the full model:

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + \gamma_{11}X_{ij}Z_j + u_{1j}X_{1ij} + u_{0j} + e_{ij}. \tag{4.7}$$

Again, if we use FML estimation, we can use the global chi-square test to formally test the improvement of fit.

If we use an exploratory procedure to arrive at a ‘good’ model, there is always the possibility that some decisions that have led to this model are based on chance. We may end up overfitting the model by following peculiarities of our specific sample, rather than characteristics of the population. If the sample is large enough, a good strategy is to split it at random into two, then to use one half for our model exploration and the other half for cross-validation of the final model. See Camstra and Boomsma (1992) for a review of several cross-validation strategies. If the sample is not large enough to permit splitting it up in an exploration and validation sample, we can apply a Bonferroni correction to the individual tests performed in the fixed part at each step.

The Bonferroni correction multiplies each $p$-value by the number of tests performed, and requires the inflated $p$-value to be significant at the usual level.[1]

At each step, we decide which regression coefficients or (co)variances to keep on the basis of the significance tests, the change in the deviance, and changes in the variance components. Specifically, if we introduce explanatory variables in step 2, we expect the lowest-level variance $\sigma_e^2$ to go down. If the composition of the groups with respect to the explanatory variables is not exactly identical for all groups, we expect the higher-level variance $\sigma_{u0}^2$ also to go down. Thus, the individual-level explanatory variables explain part of the individual and part of the group variance. The higher-level explanatory variables added in step 3 can explain only group-level variance. It is tempting to compute the analogue of a multiple correlation coefficient to indicate how much variance is actually explained at each level (see Raudenbush & Bryk, 2002). However, this 'multiple correlation' is at best an approximation, and it is quite possible for it to become smaller when we add explanatory variables, which is impossible with a real multiple correlation. This problem is taken up in section 4.5.

## 4.2 CENTERING AND STANDARDIZING EXPLANATORY VARIABLES

In ordinary multiple regression analysis, linear transformations of the explanatory variables do not change the essence of the regression estimates. If we divide an explanatory variable by two, its new regression slope equals the old one multiplied by two, the standard error is also multiplied by two, and a significance test for the regression slope gives exactly the same result. Most importantly, the proportion of unexplained residual variance, and hence the multiple correlation, does not change either. This is summed up in the statement that the multiple regression model is invariant under linear transformations; if we transform the variables, the estimated parameters change in a similar way, and it is always possible to recalculate the untransformed estimates.

In multilevel regression analysis, the model is only invariant for linear transformations if there are no random regression slopes, that is, if the slopes do not vary across the groups. To understand why this is the case, consider first a simple data set with only one explanatory variable and three groups. Figure 4.1 plots the three

[1] The usual Bonferroni correction is to keep the $p$-values, and divide the formal alpha level by the number of tests. However, if we have many tests in various steps, we end up with many different significance criteria. It is simpler to correct by appropriately inflating the $p$-values, and use one alpha criterion for all analysis steps. Both procedures are equivalent, but inflating the $p$-values makes for a simpler presentation of the results. Holm (1979) describes a more powerful variation of the Bonferroni. If $k$ tests are performed, the Holm correction would multiply the smallest $p$-value by $k$, the next smallest $p$-value by $k - 1$, and so on.

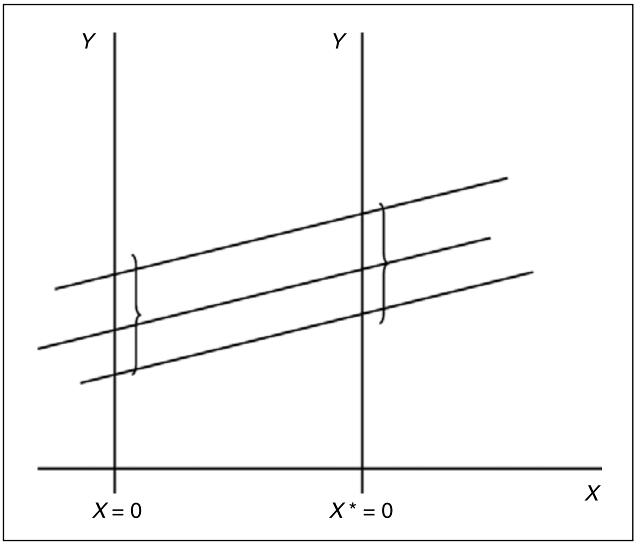


*Figure 4.1* Parallel regression lines, with shift on *X*.

regression slopes when there is no slope variance across the groups. In this situation, the slopes are parallel lines. The variance of the intercept is the variance of the slopes at the point where the slopes cut through the *Y*-axis, which is at the point where the explanatory variable *X* is equal to zero. It is clear from Figure 4.1 that, if we shift the scale of *X* to *X** by adding or subtracting a certain amount, we merely shift the location of the *Y*-axis, without altering the spread of the intercepts. In this case, the variance of the slope is clearly invariant for shifts on the *X*-axis, which we produce by adding or subtracting a constant from *X*.

The variance of the regression slopes is not invariant for such a transformation if the regression slopes vary across the groups, which is the case if we have group-level slope variance. Figure 4.2 shows the situation with three different slope coefficients. This figure shows that with random regression slopes the variance of the intercept changes when we shift the scale of the explanatory variables. It also makes clear how the intercept variance should be interpreted: it is the variance of the intercepts at the point where the explanatory variable *X* is equal to zero.

It is clear from Figure 4.2 that, if we shift the scale of *X* to *X** or *X*** by adding or subtracting a certain amount, the spread of the intercepts changes. If we shift the scale of the *X*-axis to *X**, the variation of the intercepts is considerable. If we shift the scale of the *X*-axis to *X*** and extrapolate the regression slopes, the variation of the intercepts is very small and probably not statistically significant.

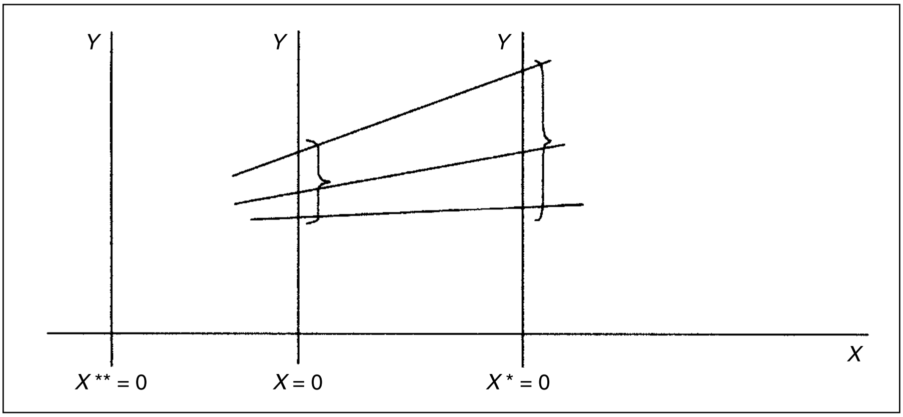


*Figure 4.2* Varying regression lines, with shifts on *X*.

In multiple regression analysis, multilevel or single-level, the intercept is interpreted as the expected value of the outcome variable, when all explanatory variables have the value zero. The problem, illustrated in Figure 4.2 by the transformation of the *X*-scale to *X***, is that in many cases 'zero' may not even be a possible value. For instance, if we have an explanatory variable 'gender' coded as 1 = male, 2 = female, zero is not in the possible score range, and as a consequence the value of the intercept is meaningless. To handle this problem, it is useful to perform a transformation of the *X*-variables that make 'zero' a legitimate, observable value.

A linear transformation that is often applied to achieve this is *centering* the explanatory variables. The usual practice is that the overall or grand mean is subtracted from all values of a variable, which is called centering on the grand mean, or grand mean centering in short. If we apply grand mean centering, we solve the problem, because now the intercept in the regression equation is always interpretable as the expected value of the outcome variable, when all explanatory variables have their mean value. Grand mean centering is most often used, but it is also possible to center on a different value, for example on the median, or on a theoretically interesting value. For example, if we have an explanatory variable 'gender' coded as 1 = male, 2 = female, the value of the mean could be 1.6, reflecting that we have 60% females in our sample and 40% males. We can center on the sample mean of 1.6, but we may prefer to center on the mean of a theoretical population with 50% males and 50% females. To accomplish this, we would center on the population mean of 1.5, which effectively leads to a code of –0.5 for male and +0.5 for female. Although we center on a value that as such does not and even cannot exist in the population, the intercept is still interpretable as the expected outcome for the average person, disregarding gender.

In multilevel modeling, centering the explanatory variables has the additional advantage that variances of the intercept and the slopes now have a clear interpretation. They are the expected variances when all explanatory variables are equal to zero, in other words, the expected variances for the 'average' subject.

Centering is also important if the multiple regression model includes interactions. For each of the two explanatory variables involved in an interaction, the interpretation of its slope is that it is the expected value of the slope when the other variable has the value zero. Again, since 'zero' may not even be a possible value, the value of the slope may not be interpretable at all. Since many multilevel regression models include cross-level interactions, this is a serious interpretation problem. When both variables in the interaction are centered on their grand mean, the problem disappears. The problem of interpreting interactions in multilevel regression models is discussed in more detail in the next section.

We will consider the issue of centering in a simple example, using the data from our example in Chapter 2, and including only pupil extraversion as an explanatory variable. We compare the estimates for pupil extraversion as a raw and as a grand mean centered variable, for a random coefficient model (varying slope for pupil extraversion). Table 4.1 also shows the estimates that result when we standardize the variable pupil extraversion. Standardization is a linear transformation that implies grand mean centering, but adds a multiplicative transformation to achieve a standard deviation of 1.

*Table 4.1* Popularity predicted by pupil extraversion, raw and centered predictor

| Model | Extraversion slope random | | |
|---|---|---|---|
| | Raw coeff. (s.e.) | Centered coeff. (s.e.) | Standardized coeff. (s.e.) |
| **Fixed part** | | | |
| Intercept | 2.46 (.20) | 5.03 (.10) | 5.03 (.10) |
| Extraversion | 0.49 (.03) | 0.49 (.03) | 0.62 (.03) |
| **Random part** | | | |
| $\sigma_e^2$ | 0.90 (.03) | 0.90 (.03) | 0.90 (.03) |
| $\sigma_{u0}^2$ | 2.94 (.58) | 0.88 (.13) | 0.88 (.13) |
| $\sigma_{u2}^2$ | 0.03 (.009) | 0.03 (.009) | 0.04 (.01) |
| **Deviance** | 5770.7 | 5770.7 | 5770.7 |

As Table 4.1 shows, grand mean centering of the variable pupil extraversion produces a different estimate for the intercept variance at the second level. The deviance remains the same, which indicates that all three random coefficient models fit the data equally well. In fact, all three models are equivalent. Equivalent models have the same fit, and produce the same residuals. The parameter estimates are not all identical, but the estimates for one model can be transformed to the estimates of the other model. Thus, grand mean centering and overall standardization do not really complicate the interpretation of the results. In addition, grand mean centering and standardization do have some advantages. One advantage is that the intercept becomes a meaningful value. The value of the higher-level intercept variance also becomes meaningful; it is the expected variance at the mean of all explanatory variables. A second advantage is that with centered explanatory variables the calculations tend to go faster, and encounter fewer convergence problems. Especially when explanatory variables vary widely in their means and variances, grand mean centering or standardization may be necessary to reach convergence, or even to be able to start the computations at all. Since grand mean centering only affects the intercept, which is often not interpreted anyway, it is preferred above standardization, which will also affect the interpretation of the regression slopes and the residual variances.

Some multilevel analysts advocate a totally different way of centering, called group mean centering. Group mean centering means that the group means are subtracted from the corresponding individual scores. Since different values are subtracted from different scores, this is not the same as centering on some overall value, such as the grand mean. This form of centering will be discussed in section 4.4.

## 4.3 INTERPRETING INTERACTIONS

Whenever there are interactions in a multiple regression analysis (whether these are a cross-level interaction in a multilevel regression analysis or an interaction in an ordinary single-level regression analysis does not matter), there are two important technical points to be made. Both stem from the methodological principle that in the presence of a significant interaction the effect of the interaction variable and the direct effects of the explanatory variables that make up that interaction must be interpreted together as a system (Aiken & West, 1991; Jaccard, Turrisi, & Wan, 1990).

The first point is that if the interaction is significant, it is best to include both direct effects in the regression too, even if they are not significant.

The second point is that in a model with an interaction effect, the regression coefficients of the simple or direct variables that make up that interaction carry a different meaning than in a model without this interaction effect. If there is an interaction, then the regression coefficient of one of the direct variables is the expected value of that regression slope when the other variable is equal to zero, and vice versa. If for one of the variables the value ‘zero’ is widely beyond the range of values that have been observed, as in age varying from 18 to 55, or if the value 'zero' is in fact impossible, as in gender coded male = 1, female = 2, the result is that the regression coefficient for the other variable has no substantive interpretation. In many such cases, if we compare different models, the regression coefficient for at least one of the variables making up the interaction will be very different from the corresponding coefficient in the model without interaction. *But this change does not mean anything*. One remedy is to take care that the value 'zero' is meaningful and actually occurs in the data. One can accomplish this by centering both explanatory variables on their grand mean.[2] After centering, the value 'zero' refers to the mean of the centered variable, and the regression coefficients do not change when the interaction is added to the model. The regression coefficient of one of the variables in an interaction can now be interpreted as the regression coefficient for individuals with an 'average' score on the other variable. If all explanatory variables are centered, the intercept is equal to the grand mean of the dependent variable.

To interpret an interaction, it is helpful to write out the regression equation for one explanatory variable for various values of the other explanatory variable. The other explanatory variables can be disregarded or included at the mean value. When both explanatory variables are continuous, we write out the regression equation for the lower-level explanatory variable, for a choice of values for the explanatory variable at the higher level. Good choices are the mean and the mean plus/minus 1 standard deviation, or the median and the 25th and 75th percentiles. A plot of these regression lines clarifies the meaning of the interaction. If one of the explanatory variables is dichotomous, we write the regression equation for the continuous variable, for both values of the dichotomous variable.

In the example we have used so far, there is a cross-level interaction between pupil extraversion and teacher experience. In the corresponding data file, pupil extraversion is measured on a 10-point scale, and the range is 1–10. Teacher experience is recorded in years, with the amount of experience ranging from 2 to 25 years. There are no pupils with a zero score on extraversion, and there are no teachers with zero experience, and this explains why adding the cross-level interaction between pupil extraversion and teacher experience to the model results in an appreciable change in the regression slope of pupil extraversion from 0.84 to 1.33. In the model without the interaction, the estimated value for the regression slope of pupil extraversion is independent from teacher experience. Therefore, it can be said to apply to the average class, with an average teacher having an amount of experience somewhere

[2] Standardizing the explanatory variables has the same effect. In this case, it is recommended not to standardize the interaction variable because that makes it difficult to compute predictions or plot interactions. Standardized regression weights for the interaction term can always be determined using equation 2.13.

in the middle between 2 and 25 years. In the model with the interaction, the pupil extraversion slope now refers to a class with a teacher who has zero years of experience. This is an extreme value, which is not even observed in the data. Following the same reasoning, we can conclude that the teacher experience slope refers to pupil extraversion = 0.

The example also includes a variable gender, coded 0 (boys) / 1 (girls). Since 'zero' is a value that does occur in the data, the interpretation of interaction effects with gender is straightforward; leaving the dummy uncentered implies that all slopes for variables interacting with gender refer to boys. This may be awkward in the interpretation, and therefore the dummy variable gender may also be centered around its grand mean or by using effect coding which codes boys = –0.5 and girls = +0.5. Centering issues do not differ for continuous and dichotomous variables (Enders & Tofighi, 2007). For a discussion of different coding schemes for categorical variables see Appendix C of this book.

The estimates for the centered explanatory variables in Table 4.2 are much more comparable across different models than the estimates for the uncentered variables (the small difference between 0.09 and 0.10 for teacher experience is because of rounding). To interpret the cross-level interaction, it helps to work out the regression equations for the effect of pupil extraversion for different values of teacher experience. Using the

*Table 4.2* Model without and with cross-level interaction

| Model | M1A: main effects | M2: with interaction | M1A: centered interaction variables | M2: centered interaction variables |
|---|---|---|---|---|
| **Fixed part** | Coeff. (s.e.) | Coeff. (s.e.) | Coeff. (s.e.) | Coeff. (s.e.) |
| Intercept | 0.74 (.20) | −1.21 (.27) | 4.39 (.06) | 4.37 (.06) |
| Gender | 1.25 (.04) | 1.24 (.04) | 1.25 (.04) | 1.24 (.04) |
| Extraversion | 0.45 (.02) | 0.80 (.04) | 0.45 (.02) | 0.45 (.02) |
| T. exp | 0.09 (.01) | 0.23 (.02) | 0.09 (.01) | 0.10 (.01) |
| Extra × T.exp | | –0.03 (.003) | | –0.025 (.002) |
| **Random part** | | | | |
| $\sigma_e^2$ | 0.55 (.02) | 0.55 (.02) | 0.55 (.02) | 0.55 (.02) |
| $\sigma_{u0}^2$ | 1.28 (.28) | 0.45 (.16) | 0.28 (.04) | 0.28 (.04) |
| $\sigma_{u2}^2$ | 0.03 (.008) | 0.005 (.004) | 0.03 (.008) | 0.005 (.004) |
| $\sigma_{u_{02}}$ | –0.18 (.05) | –0.03 (.02) | –0.01 (.02) | –0.00 (.01) |
| **Deviance** | 4812.8 | 4747.6 | 4812.8 | 4747.6 |

centered variables, the regression equation for the effect of pupil extraversion on popularity is:

$$\begin{aligned} popularity = {} & 4.368 + 1.241 \times gender + 0.451 \times extrav + 0.097 \times t.exp \\ & - 0.025 \times t.exp \times extrav. \end{aligned}$$

The average effect of a change of one scale point in extraversion is to increase the popularity by 0.451. This is the predicted value for teachers of average experience (14.2 years in the raw data set, zero in the centered data). For each year more, the effect of extraversion decreases by 0.025. So for the teachers with the most experience, 25 years in our data, the expected effect of extraversion is 0.451 – 0.025 × (25 – 14.2) = 0.18. So, for these teachers the effect of extraversion is predicted to be much smaller.

Another way to make it easier to interpret an interaction is to plot the regression slopes for one of the explanatory variables for some values of the other. The mean of pupil gender is 0.51, so we can absorb that into the intercept, giving:

$$popularity = 5.001 + 0.451 \times extrav + 0.097 \times t.exp - 0.025 \times t.exp \times extrav.$$

The centered variable pupil extraversion ranges from –4.22 to 4.79. The centered variable teacher experience ranges from –12.26 to 10.74, with a standard deviation of 6.552. We can use equation 2.12 to predict popularity, with extraversion ranging from –4.22 to 4.79 and teacher experience set at –6.552, 0, and 6.552, which are the values of 1 standard deviation below the mean, the mean, and 1 standard deviation above the mean. Figure 4.3 presents a plot of the three regression lines.

It is clear that more extraverted pupils have a higher expected popularity score, and that the difference is smaller with more experienced teachers. In general, more experienced teachers have classes with a higher average popularity. At the maximum values of teacher experience, this relationship appears to reverse, but these differences are probably not significant. If we used the uncentered scores for the plot, the scale of the $X$-axis, which represents teacher experience, would change, but the picture would not. Centering explanatory variables is especially attractive when we want to interpret the meaning of an interaction by inspecting the regression coefficients in Table 4.2. As Figure 4.3 shows, plotting the interaction over the range of observed values of the explanatory variables is an effective way to convey its meaning, even if we work with raw variables.

Interactions are sometimes interpreted in terms of moderation or a moderator effect. In Figure 4.3, we can state that the effect of pupil extraversion is moderated by teacher experience, or that the effect of teacher experience is moderated by pupil extraversion. In multilevel analysis, where the interest often lies in contextual effects, the interpretation that the effect of pupil extraversion is moderated by teacher

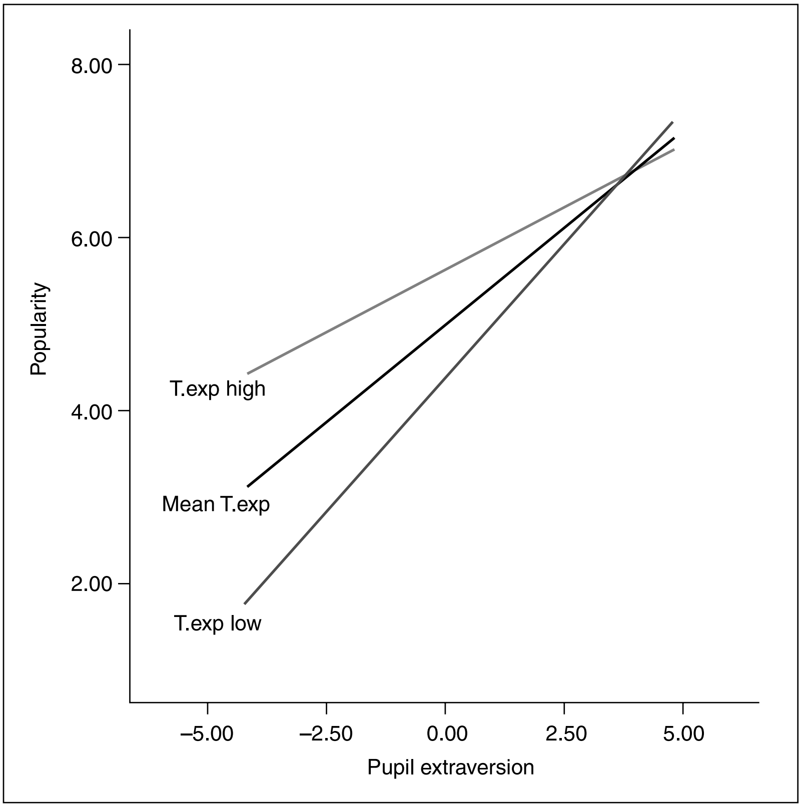


*Figure 4.3* Regression lines for popularity by extraversion, for three levels of teacher experience.

experience would in many cases be preferred. A more statistical approach is to examine the range of values of teacher experience for which the effect of pupil extraversion is significant. A simple approach to probing interactions is to test simple slopes at specific levels of the predictors. In this approach, teacher experience is centered on a range of values, to find out by trial and error where the boundary lies between a significant and a nonsignificant effect for pupil extraversion. A more general method is the Johnson-Neyman (J-N) technique, which views interactions as conditional relations in which the effect of one predictor varies with the value of another. The values of the regression coefficients and their standard errors are used to calculate the range of values on one explanaory variable for which the other variable in the interaction shows a significant effect. Bauer and Curran (2005) describe these techniques in the context of standard and multilevel regression analysis, and Preacher, Curran, and Bauer (2006) describe analytic tools to evaluate interactions by establishing confidence bands for simple slopes across the range of the moderator variable.

A final note on interactions is that in general the power of the statistical test for interactions is lower than the power for direct effects. One reason is that random slopes are estimated less reliably than random intercepts, which means that predicting slopes (using interactions with second-level variables) is more difficult than predicting intercepts (using direct effects of second-level variables; Raudenbush & Bryk, 2002). In addition, if the variables that make up an interaction are measured with some amount of measurement error, the interaction term (which is a multiplication of the two direct variables) is less reliable than the direct variables (McLelland & Judd, 1993). For these two reasons modeling random slopes is less successful than modeling random intercepts.

## 4.4 GROUP MEAN CENTERING

As mentioned in the previous section, some multilevel analysts use group mean centering, where the group's mean is subtracted from the corresponding individual scores. This is sometimes done, because it can be used to represent a specific hypothesis. For instance, in educational research there is a hypothesis called the 'frog pond effect'. This means that for the same frog the effect of being in a pond filled with big frogs is different than being in a pond filled with small frogs. In educational terms, pupils of average intelligence in a class with very smart pupils may find themselves unable to cope, and give up. Conversely, pupils of average intelligence in a class with very unintelligent pupils are relatively smart, and may become stimulated to perform really well. The frog pond hypothesis states that the effect of intelligence on school success depends on the relative standing of the pupils in their own class. A simple indicator of the pupils' relative standing can be constructed by computing the individual deviation score, by subtracting from the pupil's intelligence score the average intelligence in their class. Group mean centering is a direct translation of the frog pond mechanism in terms of the explanatory variables.

Unfortunately, group mean centering changes the meaning of the entire regression model in a complicated way. If we use grand mean centering, we get different regression slopes for variables that are involved in an interaction, and different variance components, but we have an equivalent model, with identical deviance and residual errors. Another way to describe this situation is to state that we have specified a different *parameterization* for our model, meaning that we have essentially the same model, but transformed in a way that makes it easier to interpret. Using straightforward algebra, we can transform the grand mean centered estimates back to the values we would have found by analyzing the raw scores. Group mean centering, on the contrary, is *not* a simple reparameterization of a model, but a completely different model. We will find a different deviance, and transforming the estimated parameters back to the corresponding raw score estimates is not possible. The reason is that we have subtracted not one single value, but a collection of different values from the raw scores. The technical details that govern the relations between parameter estimates using raw scores, grand mean centered scores, and group mean scores are complicated; they are discussed in detail in Kreft, de Leeuw, and Aiken (1995) and in Enders and Tofighi (2007), and in more general terms in Hofmann and Gavin (1998) and Paccagnella (2006).

As all these authors show, models become complicated when group mean centering is considered. Group mean centering of an explanatory variable implies less effective modeling than using raw scores, simply because all information concerning differences between groups is removed from that variable. It would seem reasonable to restore this information by adding the aggregated group mean again as an additional group-level explanatory variable. But this adds extra information about the group structure, which is not present in the raw scores, and therefore we will obtain a model that fits better than the raw score model. If the effect of group mean centering is analyzed in detail, it appears that group mean centering is an implicit, complicated way to change the meaning of individual- and group-level effects, including the interpretation of cross-level interactions. My recommendation is that novice users of multilevel models should apply a great deal of caution in using group mean centering. If the theory behind the modeling strongly suggests 'frog pond' or similar effects (see Enders & Tofighi, 2007; Hofmann & Gavin, 1998), group centering is an attractive approach, but the users should be keenly aware of the issues they are facing in interpretation. In the end, the choice of centering method depends on its link to the substantive research question, as grand mean centering and group mean centering address different research questions. Enders and Tofighi (2007) suggest that group mean centering (which they refer to as within-cluster centering) is most valuable (1) when the research hypothesis is about the relationship between two level 1 variables (within-group centering removes confounding with between-group effects), and (2) when a hypothesis involves interactions among level 1 variables. This includes cross-level interactions, when the research hypothesis is that the second-level predictor moderates the strength of a first-level relationship.

## 4.5 HOW MUCH VARIANCE IS EXPLAINED?

An important statistic in ordinary multiple regression analysis is the multiple correlation $R$, or the squared multiple correlation $R^2$, which is interpreted as the proportion of variance modeled by the explanatory variables. In multilevel regression analysis, the issue of modeled or explained variance is a complex one. First, there is unexplained variance at several levels to contend with. This alone makes the proportion of explained variance a more complicated concept than in single-level regression analysis. Second, if there are random slopes, the model is inherently more complex, and the concept of explained variance has no unique definition anymore. Various approaches have been proposed to indicate how well we are predicting the outcomes in a multilevel model.

A straightforward approach to examining the proportion of explained variance consists of examining the residual error variances in a sequence of models, such as the sequence proposed in Chapter 3. Table 4.3 presents for such a sequence of models the parameter estimates (regression coefficients and variance components) plus the deviance, using FML estimation. The first model is the intercept-only model. This is a useful baseline model, because it does not introduce any explanatory variables (except the constant intercept term) and decomposes the total variance of the outcome variable into two levels. Thus, the individual-level variance of the popularity scores is 1.22, the class-level variance is 0.69, and the total variance is the sum of the two: 1.91. Since there are no explanatory variables in the model, it is reasonable to interpret these variances as the error variances.

*Table 4.3* Successive models for pupil popularity data

| Model | Intercept only | Level 1 predictors | Level 2 predictors | Random coefficient | Cross-level interaction |
|---|---|---|---|---|---|
| **Fixed part** | | | | | |
| Intercept | 5.08 | 2.14 | 0.81 | 0.74 | −1.21 |
| Extraversion | | 0.44 | 0.45 | 0.45 | 0.80 |
| Gender | | 1.25 | 1.25 | 1.25 | 1.24 |
| T. exp | | | 0.09 | 0.09 | 0.23 |
| Extra × T.exp | | | | | −0.02 |
| **Random part** | | | | | |
| $\sigma_e^2$ | 1.22 | 0.59 | 0.59 | 0.55 | 0.55 |
| $\sigma_{u0}^2$ | 0.69 | 0.62 | 0.29 | 1.28 | 0.45 |
| $\sigma_{u2}^2$ | | | | 0.03 | 0.004 |
| $\sigma_{u_{02}}$ | | | | −0.18 | −0.03 |
| **Deviance** | 6327.5 | 4934.0 | 4862.3 | 4812.8 | 4747.6 |

In the first 'real' model, the pupil-level explanatory variables extraversion and gender are introduced. As a result, the first-level residual error variance goes down to 0.59, and the second-level variance goes down to 0.62. Again, it is reasonable to interpret the difference as the amount of variance explained by introducing the variables pupil gender and pupil extraversion. To calculate a statistic analogous to the multiple $R^2$, we must express this difference as a proportion of the total error variance. It appears most informative if we do this separately, level by level. For the proportion of variance explained at the first level we use (see Raudenbush & Bryk, 2002):

$$R_1^2 = \left( \frac{\sigma_{e|b}^2 - \sigma_{e|m}^2}{\sigma_{e|b}^2} \right), \tag{4.8}$$

where $\sigma_{e|b}^2$ is the lowest-level residual variance for the baseline model, which is the intercept-only model, and $\sigma_{e|m}^2$ is the lowest-level residual variance for the comparison model. For the pupil popularity data this calculates the proportion of explained variance at the pupil level for the model with pupil gender and pupil extraversion as:

$$R_1^2 = \left( \frac{1.22 - 0.59}{1.22} \right) = .52.$$

For the proportion of variance explained at the second level (see Raudenbush & Bryk, 2002) we use:

$$R_2^2 = \left( \frac{\sigma_{u0|b}^2 - \sigma_{u0|m}^2}{\sigma_{u0|b}^2} \right), \tag{4.9}$$

where $\sigma_{u0|b}^2$ is the second-level residual variance for the baseline model, which is the intercept-only model, and $\sigma_{u0|m}^2$ is the second-level residual variance for the comparison model. For the pupil popularity data this calculates the proportion of explained variance at the class level as:

$$R_2^2 = \left( \frac{0.69 - 0.62}{0.69} \right) = .10.$$

It may come as a surprise that pupil-level variables are able to explain variance at the class level. The explanation is straightforward. If the distribution of extraversion or the proportion of girls is not exactly the same in all classes, the classes do differ in their composition, and this variation can explain some of the class-level variance in average popularity between classes. In our example, the amount of variance explained by pupil extraversion and gender at the class level is small, which reflects the fact that extraversion and gender are distributed almost equally across all classes. The results could have been different; explanatory variables that are divided very selectively across the groups can often explain a fair amount of group-level variance. The interpretation would generally be that this does not reflect a real contextual effect, but rather the unequal composition of the groups.

Assessing the effect of adding the class-level explanatory variable 'teacher experience' to the model follows the same reasoning. The residual variance at the first level does not change at all. This is as it should be, because class-level variables cannot predict individual-level variation. The class-level residual variance goes down to 0.29, so the class level *R*-square becomes:

$$R_2^2 = \left( \frac{0.69 - 0.29}{0.69} \right) = .58,$$

which means that 58% of the variance at the class level is explained by the pupil gender, pupil extraversion, and teacher experience. A comparison with the previous $R_2^2 = .10$ makes clear that most of the predictive power stems from the teacher experience.

The next model is the random coefficient model, where the regression slope for pupil gender is assumed to vary across schools. In the random coefficient model, the variance of the slopes for pupil extraversion is estimated as 0.03. Since the model contains no cross-level interactions with pupil gender, this variance is not modeled, and it is analogous to the error variance of the intercept at the class level. The cross-level model includes the interaction of pupil extraversion with teacher experience, and estimates the variance for the pupil extraversion slopes as 0.004. Hence, the explained variance in these slopes is given by (see Raudenbush & Bryk, 2002):

$$R_{\beta_2}^2 = \left( \frac{\sigma_{u2|b}^2 - \sigma_{u2|m}^2}{\sigma_{u2|b}^2} \right), \tag{4.10}$$

where $\sigma_{u2|b}^2$ is the variance of the slopes for pupil extraversion in the baseline model, and $\sigma_{u2|m}^2$ is the variance of the slopes for pupil extraversion in the comparison model. For our example data, comparing the random coefficient model as a baseline model with the cross-level interaction model, we obtain (carrying one extra decimal for precision):

$$R_{extrav}^2 = \left( \frac{0.034 - 0.0047}{0.034} \right) = .86.$$

Using one explanatory variable at the school level, we can explain 86% of the variance of the pupil extraversion slopes.

All this appears straightforward, but there are two major problems. First, by using these formulas it is quite possible to arrive at the conclusion that a specific explanatory variable has a negative contribution to the explained variance. This will lead to a negative $R^2$, which is an impossible value. This is unfortunate, to say the least.

Snijders and Bosker (1994) explain why these seemingly anomalous results in fact *must* happen in some circumstances. A second problem is that, in models with random slopes, the estimated variances depend on the scale of the explanatory variables. This has been explained in section 4.2 of this chapter, in the discussion of the effects of centering and standardization. This means that the explained variance changes if we change the scale of the explanatory variables that have varying slopes. In Table 4.3 it is clear that the intercept variance changes wildly when a random slope is added to the model. Table 4.4 illustrates that using centered predictor variables produces a more stable succession of variance estimates. The regression coefficients are the same as before, except for the model with the interaction, and the deviances are all equal to the deviances using raw predictor variables.

*Table 4.4* Successive models for pupil popularity data, all predictors centered

| Model | Intercept only | Level 1 predictors | Level 2 predictors | Random coefficient | Cross-level interaction |
|---|---|---|---|---|---|
| **Fixed part** | | | | | |
| Intercept | 5.08 | 5.07 | 5.07 | 5.02 | 4.98 |
| Extraversion | | 0.44 | 0.45 | 0.45 | 0.45 |
| Gender | | 1.25 | 1.25 | 1.25 | 1.24 |
| T. exp | | | 0.09 | 0.09 | 0.09 |
| Extra × T.exp | | | | | –0.02 |
| **Random part** | | | | | |
| $\sigma_e^2$ | 1.22 | 0.59 | 0.59 | 0.55 | 0.55 |
| $\sigma_{u0}^2$ | 0.69 | 0.62 | 0.29 | 0.28 | 0.28 |
| $\sigma_{u2}^2$ | | | | 0.03 | 0.005 |
| $\sigma_{u_{02}}$ | | | | –0.01 | –0.004 |
| **Deviance** | 6327.5 | 4934.0 | 4862.3 | 4812.8 | 4747.6 |

Centering the predictor variables produces more realistic and stable variance estimates, but does not solve the problem that using equations 4.8 to 4.10 can sometimes lead to negative estimates for the explained variance. To understand how variables can make a negative contribution to the explained variance, we must investigate the effect of including an explanatory variable on the variance components. The reasoning underlying equations 4.8 to 4.10 assumes that the sample is obtained by simple random sampling at all levels. The underlying assumption is that the groups are sampled at random from a population of groups, and that the individuals are sampled at random within these groups.

Assume that we sample *N* individuals, and randomly assign them to *J* groups, all with equal group size *n*. For any variable *X* with mean $\mu$ and variance $\sigma^2$, the distribution of the group means is approximately normal with mean $\mu$ and variance:

$$\sigma_\mu^2 = \sigma^2/n \tag{4.11}$$

This is a well-known statistical theorem, which is the basis of the familiar *F*-test in the analysis of variance. In analysis of variance, we estimate the population variance $\sigma^2$ using $s_{PW}^2$, the pooled within-groups variance. A second estimate of $\sigma^2$ is given by $ns_m^2$, using equation 4.11 and filling in the observed means *m* for the population means $\mu$. This is used in the familiar *F*-test, $F = ns_m^2/s_{PW}^2$, for the null-hypothesis that there are no real differences between the groups. If there are real group differences, there is a real group-level variance $\sigma^2$ in addition to the sampling variance $\sigma_\mu^2$, and $ns_m^2$ is an estimator of $\left(\sigma^2 + \sigma_\mu^2 / n\right)$. Thus, in general, in grouped data some of the information about the population within-groups variance is in the observed between-groups variance, and the between-groups variance calculated in the sample is an upwardly biased estimator of the population between-groups variance. This also means that, even if the between-groups variance in the population is zero, the observed between-groups variance is not expected to be zero, but to be equal to $\sigma^2/n$.

As a result, for an individual-level variable sampled via a simple multilevel sampling process, we expect that it will show some between-group variability, even if there are no real group effects in the population. For such variables, the approximate *R*-square formulas defined above should do reasonably well. But in some situations we have variables that have (almost) no variation at one of the levels. This occurs when we use as the explanatory variable a group mean centered variable, from which all between-group information is removed, or the group averages, which have no variation at all at the individual level. In an implicit way this occurs when we have data with a strong selection process at one of the levels of sampling, or time series designs. For instance, if we carry out an educational experiment, we might assign pupils to classes to achieve an exactly equal gender ratio of 50% boys and 50% girls in each class. If we do this, we have no between-class variation in average gender, which is *less* than expected by simple random sampling of boys and girls. In a similar way, in many studies where we have as the lowest level a series of repeated observations at different measurement occasions, all subjects have exactly the same series of time points, because they were measured at the same occasions. Here again we have no variation of time points across subjects. In these cases, using the simple formulas given above will generally produce the result that the explanatory variable 'gender' or 'occasion' appears to explain negative variance.

Snijders and Bosker (1994) explain the problem in detail. First, let us consider a model that contains no random effects for the slopes. We could base an estimate of $\sigma_e^2$ on the pooled within-groups variance. This would be inefficient, because it ignores the information we have about $\sigma_e^2$ in the between-groups variance. Furthermore, the observed between-groups variance must be corrected for the within-groups variance to produce an accurate estimator of $\sigma_{u0}^2$. As a result, the maximum likelihood estimators of $\sigma_e^2$ and $\sigma_{u0}^2$ are a complex weighted function of the pooled within-groups and between-groups variances.

Assume that we start by estimating an intercept-only model, which gives us baseline estimates for the two variance components $\sigma_e^2$ and $\sigma_{u0}^2$. First, we introduce a 'normal' first-level explanatory variable, like pupil gender in our example. As explained above, the expected between-groups variation for such a variable is not zero, but $\sigma^2/n$. So, if this variable correlates with the outcome variable, it will reduce both the within-groups variance and the between-groups variance. The correction implicit in the ML estimators insures that both $\sigma_e^2$ and $\sigma_{u0}^2$ are reduced by the correct amount. Since $\sigma_{u0}^2$ is corrected for a 'normal' explanatory variable, it should not change, unless our explanatory variable explains some additional group-level variation as well. Now, consider what happens if we add an explanatory variable that is group mean centered, which means that all group-level information has been removed. This can reduce only the within-groups variance, and leaves the between-groups variance unchanged. The correction implicit in the ML estimator of $\sigma_{u0}^2$ will now correct for the smaller amount of within-groups variance, and as a result the estimate of the apparent between-groups variance $\sigma_{u0}^2$ will increase. Using equation 4.8, we get a negative estimate for the explained variance at the group level, which makes no sense. In ordinary multiple regression analysis such a thing cannot occur. When predictor variables are added that have more group-level variance than a random sampling process produces, the apparent within-groups variance $\sigma_{u0}^2$ can increase, which may produce a negative estimate for the explained variance at the lowest level.

With this knowledge in mind, let us look again at the formulas for explained variance. For the lowest level, we repeat the equation here:

$$R_1^2 = \left(\frac{\sigma_{e|b}^2 - \sigma_{e|m}^2}{\sigma_{e|b}^2}\right). \qquad \text{(4.8, repeated)}$$

Provided that $\sigma_e^2$ is an unbiased estimator, this formula is correct. But, as we have seen, adding group-level variables to the model may lead to incorrect estimates, because the estimation procedure does not combine the information from the two levels correctly. Snijders and Bosker (1994) propose to remedy this by replacing $\sigma_e^2$ in equation 4.8 by the sum $\sigma_e^2 + \sigma_{u0}^2$. This will use all available information about the within-groups variance in a consistent way.

The formula for the second-level explained variance is given by:

$$R_2^2 = \left(\frac{\sigma_{u0|b}^2 - \sigma_{u0|m}^2}{\sigma_{u0|b}^2}\right). \qquad \text{(4.9, repeated)}$$

Snijders and Bosker (1994) propose to replace $\sigma_{u0}^2$ in equation 4.9 by $\sigma_{u0}^2 + \sigma_e^2 \big/ n$. For unequal group sizes, the simplest solution is to replace the common group size $n$ by the average group size. A more elaborate option proposed by Snijders and Bosker (1994) is to replace $n$ by the harmonic group mean defined by $\left\{(1 \,/\, N) \sum_j (1 \,/\, n_j)\right\}^{-1}$. The ad hoc estimator proposed by Muthén (1994), given by $c = \left[N^2 - \sum n_j^2\right] \big/ \left[N(J - 1)\right]$, is probably a better replacement for the average $n$, because it is designed for similar problems in analysis of variance. Except in the case of extreme unbalance, both these replacements for $n$ are very close to the average group size $n$, so in most cases even the simple median of the group sizes $n_j$ will suffice.

The various formulas given above assume that there are no random slopes in the model. If there are, the replacements are more complicated. Assume that there are $q$ explanatory variables $Z$ with a random slope, with means $\mu_z$, between-groups covariance matrix $\sum_{\mathbf{B}}$, and pooled within-groups covariance matrix $\sum_{\mathbf{W}}$. For the first-level residual error, we replace $\sigma_e^2$ in 4.8 by:

$$\mu'_Z \sigma_{u0}^2 \mu_Z + trace\left(\sigma_{u0}^2\left(\sum\nolimits_{\mathbf{B}} + \sum\nolimits_{\mathbf{W}}\right)\right) + \sigma_e^2, \qquad \text{(4.12)}$$

and for the second-level residual error, we replace $\sigma_{u0}^2$ by:

$$\mu'_Z \sigma_{u0}^2 \mu_Z + trace\left(\sigma_{u0}^2\left(\sum\nolimits_{\mathbf{B}} + \frac{1}{n}\sum\nolimits_{\mathbf{W}}\right)\right) + \frac{1}{n}\,\sigma_e^2. \qquad \text{(4.13)}$$

The computations are straightforward, but tedious. For the development of these equations and computational details see Snijders and Bosker (1994). Snijders and Bosker (1994) advise that if there are random slopes in the model, the explained variance can still be estimated using a simplified model with only fixed effects. With grand mean centered predictor variables the $\mu$-terms are zero, and equations 4.12 and 4.13 are much simplified. If there is only one varying slope in the model, equations 4.12 and 4.13 further simplify to:

$$\sigma_{u0}^2 + \sigma_{u1}^2(\sigma_{\mathbf{B}} + \sigma_{\mathbf{W}}) + \sigma_e^2, \qquad \text{(4.12a)}$$

and

$$\sigma_{u0}^2 + \sigma_{u1}^2\left(\sigma_{\mathbf{B}} + \frac{1}{n}\sigma_{\mathbf{W}}\right) + \frac{1}{n}\,\sigma_e^2. \qquad \text{(4.13a)}$$

Table 4.5 lists the models presented in Table 4.4, with the explained variance as calculated using the approximations in equations 4.8 and 4.9, and with the explained variance as calculated using the Snijders and Bosker (1994) corrections.

*Table 4.5* 'Explained variance' for pupil popularity models in Table 4.4

| Model | Intercept only | Level 1 predictors | Level 2 predictors | Random coefficient | Cross-level interaction |
|---|---|---|---|---|---|
| $R_1^2$ (approx.) | .00 | .52 | .52 | .55 | .55 |
| $R_2^2$ (approx.) | .00 | .10 | .58 | .59 | .59 |
| $R_{\text{extr.}}^2$ (approx.) | – | – | – | .00 | .83[a] |
| $R_1^2$ ($S$ & $B$) | .00 | .37 | .54 | .54 | .54 |
| $R_2^2$ ($S$ & $B$) | .00 | .14 | .57 | .57 | .57 |

[a] Calculated carrying only decimals reported in Table 4.4.

The explained variance at the first level differs appreciably in the different approaches. There is a difference in interpretation as well. The approximate $R^2$ attempts to indicate the explained variance as a proportion of the first-level variance only. The Snijders and Bosker $R^2$ attempts to indicate the explained variance as a proportion of the total variance, because in principle first-level variables can explain all variation, including the variation at the second level. Note that the approximate $R^2$ appears to increase when a random effect is added to the model. This increase is spurious, because we cannot actually observe the random $u$-terms. Thus, although the approximate approach can be used to compute $R^2$ for models including random effects, the results must be interpreted with care. The Snijders and Bosker correction removes this spurious increase. Hence, for the approximate procedure, the explained intercept variance is based on a comparison of the null model and the model without random slopes.

Snijders and Bosker (1994) do not address the problem of defining explained variance in regression slopes, or explained variance in models with more than two levels. It is clear that the problems just described are also present at any higher levels, with more complex equations as the result. Since the intercept is simply the regression slope associated with a constant with value 1, the complications in interpreting the intercept variance will arise again when we consider the variance of the regression slopes. In addition, Snijders and Bosker's approach does not solve the problem that with random slopes the residual variances depend on the scale of the explanatory variables. There is no current solution for this.

When the multilevel sampling process is close to two-stage simple random sampling, the formulas given as 4.8 to 4.10 should give reasonable approximations. As always, when the interest is in the size of variance components, it appears prudent to center all explanatory variables that have varying slopes on their grand mean. Given that the estimates of the variance components depend on the explanatory variables, this at least insures that we get variance estimates for values of the explanatory variables that actually exist, and reflect the average sampling unit. This would apply to both the approximate and the Snijders and Bosker approach to estimating explained variance.

# 5

# Analyzing Longitudinal Data

Longitudinal data, or repeated measures data, can be viewed as multilevel data with repeated measurements nested within individuals. In its simplest form, this leads to a two-level model, with the series of repeated measures at the lowest level, and the individual subjects at the highest level. Longitudinal measures can be taken at fixed or at varying occasions. Multilevel analysis for longitudinal data can handle both situations. Since multilevel modeling does not require balanced data, it is not a problem if the number of available measurements is not the same for all individuals. This is an important benefit if there is panel dropout, or other forms of missing measurements within individuals. Since longitudinal data collected at fixed occasions is the simplest situation, this chapter starts with fixed occasions, and discusses varying occasions later.

If the data are collected to analyze individual change over time, the constructs under study must be measured on a comparable scale at each occasion (see Plewis, 1996; Taris, 2000). When the time span is short, this does not pose complicated problems. For instance, Tate and Hokanson (1993) report on a longitudinal study where the scores of students on the Beck Depression Scale were collected at three occasions during the academic year. In such an application, especially when a well-validated measurement instrument is used, we may assume that the research instrument remains constant for the duration of the study. On the other hand, in a study that examines improvements in reading skill in school children from ages 5 to 12, it is clear that we cannot use the same instrument to measure reading skill at such different age levels. Here, we must make sure that the different measurement instruments are calibrated, meaning that a specific score has the same psychometric meaning at all age levels, independent of the actual reading test that is used. The issues are the same as the issues in cross-cultural comparison (see Bechger, van Schooten, de Glopper, & Hox, 1998). Another requirement is that there is sufficient time between the measurements that memory effects are not a problem. In some applications, this may not be the case. For instance, if data are collected that are closely spaced in time, we may expect considerable correlation between measurements collected at occasions that are close together, partly because of memory effects. These effects should then be included in the model, which leads to models with correlated errors. Formulating multilevel models for such situations can be quite complex. Some multilevel software has built-in provisions for modeling correlated errors. These are discussed in the last part of this chapter.

The models discussed in this chapter are all models for data that have repeated measures on individuals over time. Within the framework of multilevel modeling, we can also analyze data where the repeated measures are on higher levels, for example data where we follow the same set of schools over a number of years, with of course in each year a different set of pupils. Models for such data are similar to the models discussed in this chapter. Such repeated cross-sectional data are discussed by DiPrete and Grusky (1990) and Raudenbush and Chan (1993). Multilevel analysis models for longitudinal data are discussed in detail by Hedeker and Gibbons (2006) and by Singer and Willet (2003). Latent curve analysis using structural equation modeling is discussed by Duncan, Duncan, and Strycker (2006) and by Bollen and Curran (2006). The structural equation approach to latent curve analysis is treated in this book in Chapter 16.

Multilevel analysis of repeated measures is often applied to data from large-scale panel surveys. In addition, it can also be a valuable analysis tool in a variety of experimental designs. If we have a pretest–posttest design, the usual analysis is an analysis of covariance (ANCOVA) with the experimental and control groups as the factor and the pretest as the covariate. In the multilevel framework, we analyze the slopes of the change over time, using an experimental group/control group dummy variable to predict differences in the slopes. If we have just a pretest–posttest design this does not offer much more than the usual analysis of covariance. However, in the multilevel framework it makes sense to add more measurement occasions between the pretest and the posttest. Willett (1989) and Maxwell (1998) show that the power of the test for differences between the experimental and the control groups can be increased dramatically by adding only a few additional waves of data collection. There is also an advantage on ANCOVA if there is dropout, especially if this is not completely random. Multilevel analysis of repeated measures can include incomplete cases, which is a major advantage when incomplete data indeed occur.

## 5.1 FIXED AND VARYING OCCASIONS

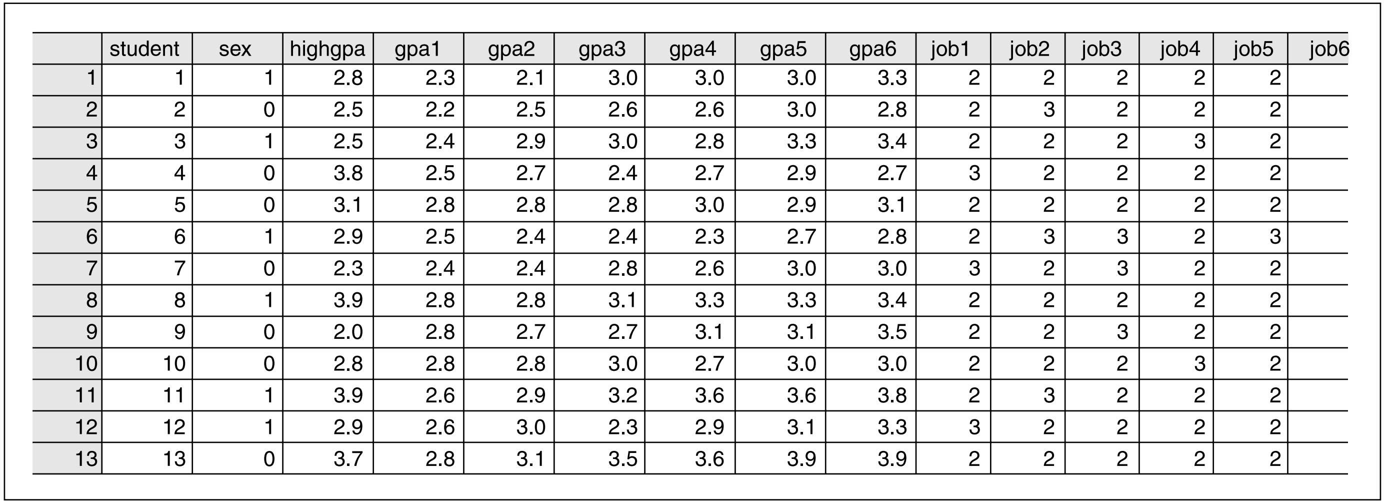
It is useful to distinguish between repeated measures that are collected at fixed or varying occasions. If the measurements are taken at fixed occasions, all individuals provide measurements at the same set of occasions, usually regularly spaced, such as once every year. When occasions are varying, we have a different set of measures taken at different points in time for different individuals. Such data occur, for instance, in growth studies, where physical or psychological characteristics are studied for a set of individuals at different moments in their development. The data collection could be at fixed moments in the year, but the individuals would have different ages at that moment. Alternatively, the original design is a fixed occasion design, but because of planning problems, the data collection does not take place at the intended moments. For a multilevel analysis of the resulting data, the difference between fixed and varying occasions is not very important. For fixed occasion designs, especially when the occasions are regularly spaced and when there are no missing data, repeated measures analysis of variance (ANOVA) is a viable alternative for multilevel analysis. A comparison of the ANOVA approach and multilevel analysis is given in section 5.2. Another possibility in such designs is latent curve analysis, also known as latent growth curve analysis. This is a structural equation model (see Duncan et al., 2006; Singer & Willett 2003) that models a repeated measures polynomial analysis of variance. Latent growth curve models are treated in Chapter 16. Multilevel models for longitudinal data are discussed by, among others, Bryk and Raudenbush (1987, Raudenbush & Bryk, 2002) and Goldstein (2003); for introductory articles see Snijders (1996), Cnaan, Laird, and Slasor (1997), and Hedeker and Gibbons (2006).

## 5.2 EXAMPLE WITH FIXED OCCASIONS

The example data are a longitudinal data set from 200 college students. The students' grade point average (GPA) has been recorded for six successive semesters. At the same time, it was recorded whether the student held a job in that semester, and for how many hours. This is recorded in a variable 'job' (= hours worked). In this example, we also use the student-level variables high school GPA and gender (0 = male, 1 = female), which of course remain constant for each student across the six measurement occasions.

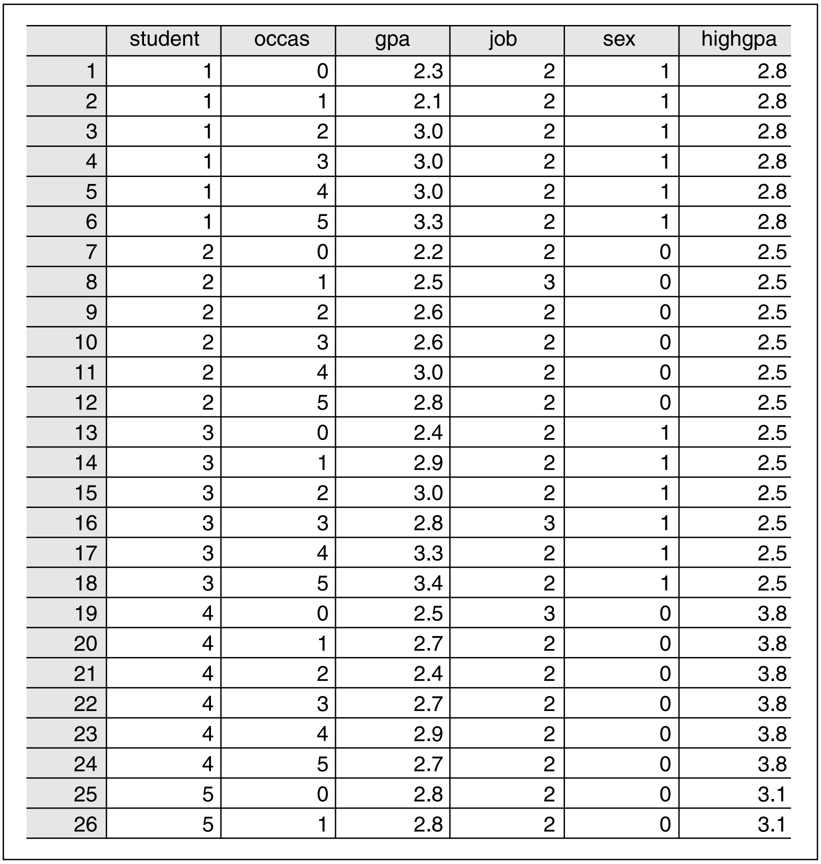
In a statistical package such as SPSS or SAS, such data are typically stored with the students defining the cases, and the repeated measurements as a multivariate set of variables, such as GPA1, GPA2, . . . , GPA6, and JOB1, JOB2, . . . , JOB6. For example, in SPSS the data structure would be as shown in Figure 5.1.

The data structure for a multilevel analysis of these data is generally different, depending on the specific program that is used. Most multilevel software requires that the data are structured with the measurement occasions defining the lowest level, and student-level variables repeated over the cases. Figure 5.2 presents the GPA data in this format, where each row in the data set represents a separate occasion, with repeated measurements resulting in six rows for each student. This data format is sometimes referred to as a 'long' (or 'stacked') data set, and the regular format in Figure 5.1 is referred to as a 'wide' data set (see Chapter 10 on multivariate multilevel analysis). Although Figures 5.1 and 5.2 do not include missing data, missing occasions simply result in students with less than the full set of six occasions in the data file. As a result, missing occasions are very simple to handle in a multilevel model. Note that the measurement occasions are numbered 0, . . . , 5 instead of 1, . . . , 6. This ensures that 'zero' is part of the range of possible values. For the data in Figure 5.2, the intercept



| | student | sex | highgpa | gpa1 | gpa2 | gpa3 | gpa4 | gpa5 | gpa6 | job1 | job2 | job3 | job4 | job5 | job6 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 2.8 | 2.3 | 2.1 | 3.0 | 3.0 | 3.0 | 3.3 | 2 | 2 | 2 | 2 | 2 | |
| 2 | 2 | 0 | 2.5 | 2.2 | 2.5 | 2.6 | 2.6 | 3.0 | 2.8 | 2 | 3 | 2 | 2 | 2 | |
| 3 | 3 | 1 | 2.5 | 2.4 | 2.9 | 3.0 | 2.8 | 3.3 | 3.4 | 2 | 2 | 2 | 3 | 2 | |
| 4 | 4 | 0 | 3.8 | 2.5 | 2.7 | 2.4 | 2.7 | 2.9 | 2.7 | 3 | 2 | 2 | 2 | 2 | |
| 5 | 5 | 0 | 3.1 | 2.8 | 2.8 | 2.8 | 3.0 | 2.9 | 3.1 | 2 | 2 | 2 | 2 | 2 | |
| 6 | 6 | 1 | 2.9 | 2.5 | 2.4 | 2.4 | 2.3 | 2.7 | 2.8 | 2 | 3 | 3 | 2 | 3 | |
| 7 | 7 | 0 | 2.3 | 2.4 | 2.4 | 2.8 | 2.6 | 3.0 | 3.0 | 3 | 2 | 3 | 2 | 2 | |
| 8 | 8 | 1 | 3.9 | 2.8 | 2.8 | 3.1 | 3.3 | 3.3 | 3.4 | 2 | 2 | 2 | 2 | 2 | |
| 9 | 9 | 0 | 2.0 | 2.8 | 2.7 | 2.7 | 3.1 | 3.1 | 3.5 | 2 | 2 | 3 | 2 | 2 | |
| 10 | 10 | 0 | 2.8 | 2.8 | 2.8 | 3.0 | 2.7 | 3.0 | 3.0 | 2 | 2 | 2 | 3 | 2 | |
| 11 | 11 | 1 | 3.9 | 2.6 | 2.9 | 3.2 | 3.6 | 3.6 | 3.8 | 2 | 3 | 2 | 2 | 2 | |
| 12 | 12 | 1 | 2.9 | 2.6 | 3.0 | 2.3 | 2.9 | 3.1 | 3.3 | 3 | 2 | 2 | 2 | 2 | |
| 13 | 13 | 0 | 3.7 | 2.8 | 3.1 | 3.5 | 3.6 | 3.9 | 3.9 | 2 | 2 | 2 | 2 | 2 | |

*Figure 5.1* Repeated measures data structure in SPSS.



| | student | occas | gpa | job | sex | highgpa |
|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 2.3 | 2 | 1 | 2.8 |
| 2 | 1 | 1 | 2.1 | 2 | 1 | 2.8 |
| 3 | 1 | 2 | 3.0 | 2 | 1 | 2.8 |
| 4 | 1 | 3 | 3.0 | 2 | 1 | 2.8 |
| 5 | 1 | 4 | 3.0 | 2 | 1 | 2.8 |
| 6 | 1 | 5 | 3.3 | 2 | 1 | 2.8 |
| 7 | 2 | 0 | 2.2 | 2 | 0 | 2.5 |
| 8 | 2 | 1 | 2.5 | 3 | 0 | 2.5 |
| 9 | 2 | 2 | 2.6 | 2 | 0 | 2.5 |
| 10 | 2 | 3 | 2.6 | 2 | 0 | 2.5 |
| 11 | 2 | 4 | 3.0 | 2 | 0 | 2.5 |
| 12 | 2 | 5 | 2.8 | 2 | 0 | 2.5 |
| 13 | 3 | 0 | 2.4 | 2 | 1 | 2.5 |
| 14 | 3 | 1 | 2.9 | 2 | 1 | 2.5 |
| 15 | 3 | 2 | 3.0 | 2 | 1 | 2.5 |
| 16 | 3 | 3 | 2.8 | 3 | 1 | 2.5 |
| 17 | 3 | 4 | 3.3 | 2 | 1 | 2.5 |
| 18 | 3 | 5 | 3.4 | 2 | 1 | 2.5 |
| 19 | 4 | 0 | 2.5 | 3 | 0 | 3.8 |
| 20 | 4 | 1 | 2.7 | 2 | 0 | 3.8 |
| 21 | 4 | 2 | 2.4 | 2 | 0 | 3.8 |
| 22 | 4 | 3 | 2.7 | 2 | 0 | 3.8 |
| 23 | 4 | 4 | 2.9 | 2 | 0 | 3.8 |
| 24 | 4 | 5 | 2.7 | 2 | 0 | 3.8 |
| 25 | 5 | 0 | 2.8 | 2 | 0 | 3.1 |
| 26 | 5 | 1 | 2.8 | 2 | 0 | 3.1 |

*Figure 5.2* Repeated measures data structure for multilevel analysis.

can be interpreted as the starting value at the first measurement occasion, and the second-level variance is the variance at the first measurement occasion. Other coding schemes for the measurement occasions are possible, and will be discussed later in this chapter.

The multilevel regression model for longitudinal data is a straightforward application of the multilevel regression model described in Chapter 2. It can also be written as a sequence of models for each level. At the lowest, the repeated measures level, we have:

$$Y_{ti} = \pi_{0i} + \pi_{1i}T_{ti} + \pi_{2i}X_{ti} + e_{ti}. \tag{5.1}$$

In repeated measures applications, the coefficients at the lowest level are often indicated by the Greek letter $\pi$. This has the advantage that the subject-level coefficients, which in repeated measures are at the second level, can be represented by the usual Greek letter $\beta$, and so on. In equation 5.1, $Y_{ti}$ is the response variable of individual $i$ measured at measurement occasion $t$, $T$ is the time variable that indicates the measurement occasion, and $X_{ti}$ is a *time-varying covariate*. For example, $Y_{ti}$ could be the GPA of a student at measurement occasion $t$, $T_{ti}$ indicates the occasion at which the GPA is measured, and $X_{ti}$ the job status of the student at time $t$. Student characteristics, such as gender, are *time-invariant covariates*, which enter the equation at the second level:

$$\begin{aligned}\pi_{0i} &= \beta_{00} + \beta_{01}Z_i + u_{0i}\\ \pi_{1i} &= \beta_{10} + \beta_{11}Z_i + u_{1i}\\ \pi_{2i} &= \beta_{20} + \beta_{21}Z_i + u_{2i}.\end{aligned} \tag{5.2}$$

By substitution, we get the single equation model:

$$\begin{aligned}Y_{ti} = {} & \beta_{00} + \beta_{10}T_{ti} + \beta_{20}X_{ti} + \beta_{01}Z_i + \beta_{11}T_{ti}Z_i + \beta_{21}X_{ti}Z_i\\ & + u_{1i}T_{ti} + u_{2i}X_{ti} + u_{0i} + e_{ti}.\end{aligned} \tag{5.3}$$

Using variable labels instead of letters, the equation for our GPA example becomes:

$$\begin{aligned}Y_{ti} = {} & \beta_{00} + \beta_{10}Occasion_{ti} + \beta_{20}Job_{ti} + \beta_{01}Sex_i\\ & + \beta_{11}Occasion_{ti}Sex_i + \beta_{21}Job_{ti}Sex_i\\ & + u_{1i}Occasion_{ti} + u_{2i}Job_{ti} + u_{0i} + e_{ti}.\end{aligned} \tag{5.4}$$

In longitudinal research, we sometimes have repeated measurements of individuals, who are all measured together on a small number of fixed occasions. This is typically the case with experimental designs involving repeated measures and panel research. If we simply want to test the null-hypothesis that the means are equal for all occasions, we can use repeated measures analysis of variance. If we use repeated measures univariate analysis of variance (Stevens, 2009, p. 420), we must assume *sphericity*. Sphericity means that there are complex restrictions on the variances and covariances between the repeated measures; for details see Stevens (2009, Chapter 13). A specific form of sphericity, which is easily understood, is *compound symmetry*, sometimes referred to as *uniformity*. Compound symmetry requires that all population variances of the repeated measures are equal, and that all population covariances of the repeated measures are equal. If sphericity is not met, the *F*-ratio used in analysis of variance is positively biased, and we reject the null-hypothesis too often. A different approach is to specify the repeated measures as observations on a multivariate response vector and use multivariate analysis of variance (MANOVA). This does not require sphericity, and is considered the preferred approach if analysis of variance is used on repeated measures (O'Brien & Kaiser, 1985; Stevens, 2009). However, the multivariate test is more complicated, because it is based on a transformation of the repeated measures, and what is tested are actually contrasts among the repeated measures.

A MANOVA analysis of the example data using the general linear model in SPSS (SPSS Inc., 1997) cannot easily incorporate a time-varying covariate such as job status. But MANOVA can be used to test the trend over time of the repeated GPA measures by specifying polynomial contrasts for the measurement occasions, and to test the fixed effects of gender and high school GPA. Gender is a dichotomous variable, which is entered as a factor, and high school GPA is a continuous variable that is entered as a covariate. Table 5.1 presents the results of the traditional significance tests.

*Table 5.1* MANOVA significance tests on GPA example data

| Effect tested | *F* | *df* | *p* |
|---|---|---|---|
| Occasion | 3.58 | 5/193 | .004 |
| Occ. (linear) | 8.93 | 1/197 | .003 |
| Occ. × HighGPA | 0.87 | 5/193 | .505 |
| Occ. × Gender | 1.42 | 5/193 | .220 |
| HighGPA | 9.16 | 1/197 | .003 |
| Gender | 18.37 | 1/197 | .000 |

The MANOVA results indicate that there is a significant linear trend for the GPA measures. Both gender and high school GPA have significant effects. The higher polynomial trends, which are not in the table, are not significant, and the interactions between measurement occasion and high school GPA and gender are not significant. Table 5.2 presents the GPA means at the different measurement occasions, rounded to one decimal, for all six occasions, for male and female students.

*Table 5.2* GPA means at six occasions, for male and female students

| Occasion | 1 | 2 | 3 | 4 | 5 | 6 | Total |
|---|---|---|---|---|---|---|---|
| Male | 2.6 | 2.7 | 2.7 | 2.8 | 2.9 | 3.0 | 2.8 |
| Female | 2.6 | 2.8 | 2.9 | 3.0 | 3.1 | 3.2 | 2.9 |
| All students | 2.6 | 2.7 | 2.8 | 2.9 | 3.0 | 3.1 | 2.9 |

Table 5.2 makes clear that there is a linear upward trend of about 0.1 for each successive GPA measurement. Female students have a GPA that is consistently higher than the GPA of the male students. Finally, the SPSS output also contains the regression coefficients for the gender and high school GPA at the six occasions; these coefficients (not given in the table) are different for each predicted occasion, but both generally positive, indicating that female students do better than males on each occasion, and that students who have a high GPA in high school have a relatively high GPA in college at each measurement occasion.

In the multilevel regression model, the development over time is often modeled by a linear or polynomial regression equation, which may have different regression coefficients for different individuals. Thus, each individual can have their own regression curve, specified by the individual regression coefficients that in turn may depend on individual attributes. Quadratic and higher functions can be used to model nonlinear dependencies on time, and both time-varying and subject-level covariates can be added to the model. Although the measurement occasions will usually be thought of as occasion 1, 2, and so on, it is useful to code the measurement occasions $T$ as $t = 0$, 1, 2, 3, 4, 5. As a result, the intercept can be interpreted as the expected outcome on the first occasion. Using measurement occasions $t = 1, 2, 3, 4, 5, 6$ would be equivalent, but more difficult to interpret, because the value zero is not in the range of observed measurement occasions.[1] If the explanatory variable is not successive measurement occasions but, for instance, calendar age, setting the first observation to zero is not the best solution. In that case, it is usual to center on the mean or median age, or on a rounded-off value close to the mean or median.

Before we start the analysis, we examine the distribution of the outcome variable GPA in the disaggregated data file with 200 × 6 = 1200 observations. The histogram with embedded best-fitting normal curve is in Figure 5.3. The distribution appears quite normal, so we proceed with the analysis.

Table 5.3 presents the results of a multilevel analysis of these longitudinal data. Model 1 is a model that contains only an intercept term and variances at the occasion and the subject level. The intercept of 2.87 in this model is simply the average GPA across all individuals and occasions. The intercept-only model estimates the repeated measures (level 1) variance as 0.098, and the subject-level (level 2) variance as 0.057 (because these numbers are so small, they are given to 3 decimal places). This estimates the total GPA variance as 0.155. Using equation 2.9, the intraclass correlation or the proportion variance at the subject level is estimated as $\rho = 0.057/0.155 = .37$. About one-third of the variance of the six GPA measures is variance between individuals, and about two-thirds is variance within individuals across time.

[1] The importance of centering explanatory variables on their overall mean or a similar value is discussed in Chapter 4.

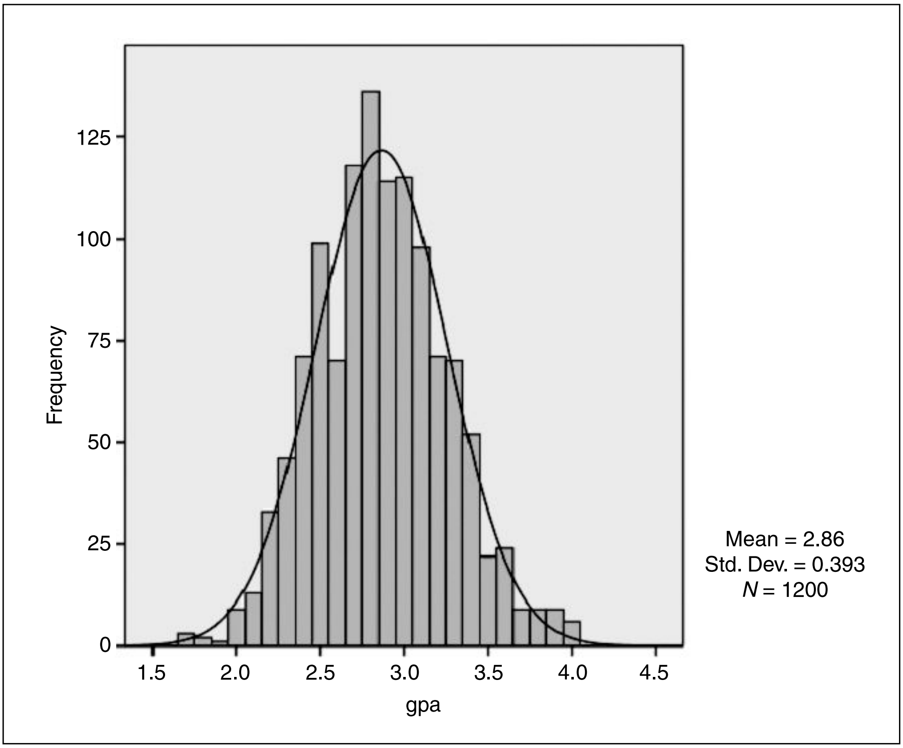


*Figure 5.3* Histogram of GPA values in disaggregated data file.

In model 2, the time variable is added as a linear predictor with the same coefficient for all subjects. The model predicts a value of 2.60 at the first occasion, which increases by 0.11 on each succeeding occasion. Just as in the MANOVA analysis, adding higher order polynomial trends for time to the model does not improve prediction. Model 3 adds the time-varying covariate job status to the model. The effect of job status is clearly significant; the more hours are worked, the lower the GPA. Model 4 adds the subject-level (time-invariant) predictors high school GPA and sex. Both effects are significant; high school GPA correlates with average GPA in college, and female students perform better than male students.

In all models in Table 5.3, the Wald test indicates that the subject-level (second-level) variance is significant. The more accurate test using the difference of the deviance in a model with and a model without the second-level variance term confirms this for all models in the table (results not reported here).

If we compare the variance components of model 1 and model 2, we see that entering the measurement occasion variable decreases the occasion-level variance

*Table 5.3* Results of multilevel analysis of GPA, fixed effects

| Model | M1: null model | M2: + occas. | M3: + job stat | M4: + high school GPA, gender |
|---|---|---|---|---|
| **Fixed part** | Coeff. (s.e.) | Coeff. (s.e.) | Coeff. (s.e.) | Coeff. (s.e.) |
| Intercept | 2.87 (.02) | 2.60 (.02) | 2.97 (.04) | 2.64 (.10) |
| Occasion | | 0.11 (.004) | 0.10 (.003) | 0.10 (.003) |
| Job status | | | –0.17 (.02) | –0.17 (.02) |
| GPA highschl | | | | 0.08 (.03) |
| Gender | | | | 0.15 (.03) |
| **Random part** | | | | |
| $\sigma^2_e$ | 0.098 (.004) | 0.058 (.025) | 0.055 (.002) | 0.055 (.002) |
| $\sigma^2_{u0}$ | 0.057 (.007) | 0.063 (.007) | 0.052 (.006) | 0.045 (.01) |
| Deviance | 913.5 | 393.6 | 308.4 | 282.8 |
| AIC | 919.5 | 401.6 | 318.4 | 296.8 |
| BIC | 934.7 | 422.0 | 343.8 | 332.4 |

considerably, while increasing the subject-level variance by as much as 11%. If the usual formula is used to estimate the second-level variance explained by the measurement occasion variable, we arrive at a negative value for the amount of explained variance. This is odd, but it is in fact typical for multilevel analysis of repeated measures. The occurrence of negative estimates for the explained variance makes it impossible to use the residual error variance of the intercept-only model as a benchmark, and to examine how much this goes down when explanatory variables are added to the model.

The reason for this apparent anomaly is, as is discussed in detail in Chapter 4, that the 'amount of variance explained at a specific level' is not a simple concept in multilevel models (see Snijders & Bosker, 1994). The problem arises because the statistical model behind multilevel models is a hierarchical sampling model: groups are sampled at the higher level, and at the lower level individuals are sampled within groups. This sampling process creates some variability in all variables between the groups, even if there are in fact no real group differences. In a time series design, the lowest level is a series of measurement occasions. In many cases, the data collection design is set up to make sure that the repeated measurements are evenly spaced and the data are collected at the same time for all individuals in the sample. Therefore, the variability between subjects in the measurement occasion variable is usually *much* higher than the hierarchical sampling model assumes. Consequently, the intercept-only model overestimates the variance at the occasion level, and underestimates the variance at the subject level. Model 2 uses the measurement occasion variable to model the occasion-level variance in the dependent variable GPA. Conditional on this effect, the variances estimated at the measurement occasions and at the subject level are much more realistic.

Chapter 4 in this book describes procedures based on Snijders and Bosker (1994) to correct the problem. A simple approximation is to use as a baseline model for the 'explained variance' a model that includes the measurement occasion in an appropriate manner. Whether this is linear or needs to be some polynomial must be determined by preliminary analyses. In our example, a linear trend for measurement occasion suffices. Using M2 in Table 5.3 as the baseline, we calculate that job status explains (0.058 – 0.055)/0.058 = 0.052 or 5.2% of the variance, indicating that in semesters that they work more hours off campus, students tend to have a lower grade. The time-varying predictor job status explains a further (0.063 – 0.052/0.063) = 0.175 or 17.5% of the variance between students; apparently students differ in how many hours they work in an off-campus job. Hence, although job status is a time-varying predictor, it explains more variation between different subjects in the same semester than within the same subjects from one semester to the next. The pupil-level variables gender and high school GPA explain an additional 11.5% of the between-students variance.

The models presented in Table 5.3 all assume that the rate of change is the same for all individuals. In the models presented in Table 5.4, the regression coefficient of the measurement occasion variable is assumed to vary across individuals.

In model 5 in Table 5.4, the slope of the measurement occasion variable is allowed to vary across individuals. The Wald test for the variance of the slopes for occasion is significant, $Z = 6.02$ (calculated carrying more decimal values than reported in Table 5.4). The deviance difference test (comparing model 5 to the same model without the subject-level variance) produces a chi-square of 109.62. With one degree of freedom, this translates to $Z = 10.47$, which demonstrates again that for variances the deviance difference test is generally more powerful than the Wald test.

The variance components for the intercept and the regression slope for the time variable are both significant. The significant intercept variance of 0.038 means that individuals have different initial states, and the significant slope variance of 0.004 means that individuals also have different rates of change. In model 6, the interaction of the occasion variable with the subject level predictor gender is added to the model. The interaction is significant, but including it does not decrease the slope variance for the time variable (actually, carrying all decimals in the output leads to a decrease in slope variance of 0.00022).

The variance component of 0.004 for the slopes of the occasion variable does not seem large. However, multilevel models assume a normal distribution for these slopes (or, equivalently, for the slope residuals $u_1$), for which the standard deviation is

*Table 5.4* Results of multilevel analysis of GPA, varying effects for occasion

| Model | M5: + occasion random | M6: + cross-level interaction | M6: standardized |
|---|---|---|---|
| **Fixed part** | Coeff. (s.e.) | Coeff. (s.e.) | |
| Intercept | 2.56 (.10) | 2.58 (.09) | |
| Occasion | 0.10 (.006) | 0.09 (.01) | 0.38 |
| Job status | –0.13 (.02) | –0.13 (.02) | –0.14 |
| GPA highschl | 0.09 (.03) | 0.09 (.03) | 0.13 |
| Gender | 0.12 (.03) | 0.08 (.03) | 0.10 |
| Occas*Gender | | 0.03 (.01) | 0.13 |
| **Random part** | | | |
| $\sigma^2_e$ | 0.042 (.002) | 0.042 (.002) | |
| $\sigma^2_{u0}$ | 0.038 (.006) | 0.038 (.01) | |
| $\sigma^2_{u1}$ | 0.004 (.001) | 0.004 (.001) | |
| $\sigma_{u01}$ | –0.002 (.002) | –0.002 (.001) | |
| $r_{u01}$ | –.21 | –.19 | |
| Deviance | 170.1 | 163.0 | |
| AIC | 188.1 | 183.0 | |
| BIC | 233.93 | 233.87 | |

estimated in models 5 and 6 as $\sqrt{0.004} = 0.063$. Compared to the value of 0.10 for the average time slope in model 5, this is not very small. There is substantial variation among the time slopes, which is not modeled well by the available student variables.

In both model 5 and model 6 there is a small negative covariance $\sigma_{u01}$ between the initial status and the growth rate; students who start with a relatively low value of their GPA increase their GPA faster than the other students. It is easier to interpret this covariance if it is presented as a correlation between the intercept and slope residuals. Note that the correlation $r_{u01}$ between the intercept and slope is slightly different in models 5 and 6; the covariances seem equal because of rounding. In a model without other predictors except the time variable, this correlation can be interpreted as an ordinary correlation, but in models 5 and 6 it is a partial correlation, conditional on the predictors in the model.

When the fit indices AIC and BIC are inspected, they both indicate model 6 as the best model. Since the slope variation is small but not negligible, and since the cross-level interaction is also significant, we decide to keep model 6.

To facilitate interpretation, standardized regression coefficients are calculated for the last model (model 6) in Table 5.4 using equation 2.13. The standardized regression coefficients indicate that the change over time is the largest effect. The standardized results also suggest that the interaction effect is more important than the unstandardized analyses indicate. To investigate this further, we can construct the regression equation of the time variable separately for both male and female students. Since gender in this example is coded 0 (male) and 1 (female), including the interaction changes the value of the regression coefficient for the time trend. As discussed in Chapter 4, this regression coefficient now reflects the expected time effect for respondents with value zero on the gender variable. Thus, the regression coefficient of 0.09 for occasion in the final model refers to the male students. For female students the interaction term is added, so their regression coefficient equals 0.09 + 0.03 = 0.12.

Figure 5.4 presents a plot of these regression lines. The expected difference between male and female students, which is 0.08 in the first semester, increases to 0.11 in the second semester. In the sixth semester, the difference has grown to 0.23.

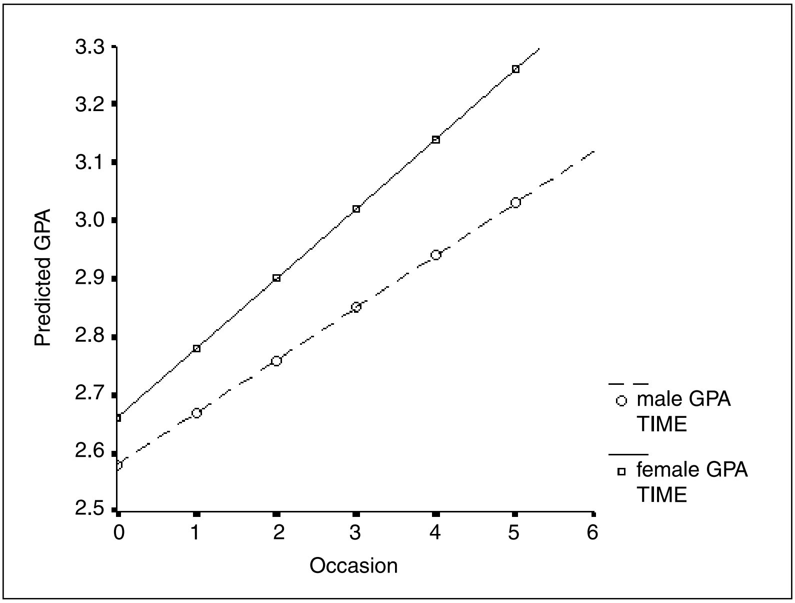


*Figure 5.4* Regression lines for occasion, separate for male and female students.

Since the measurement occasion variable is coded in such a way that the first occasion is coded as zero, the negative correlation between the intercepts and slopes refers to the situation on the first measurement. As is explained in section 4.2 of

Chapter 4, the estimates of the variance components in the random part can change if the scale of the time variable is changed. In many models, this is not a real problem, because the interest is mostly in estimation and interpretation of the regression coefficients in the fixed part of the model. In repeated measures analysis, the correlation between the intercepts and the slopes of the time variable is often an interesting parameter, to be interpreted along with the regression coefficients. In this case, it is important to realize that this correlation is *not* invariant; it changes if the scale of the time variable is changed. In fact, one can show that by using extremely different scalings for the time variable, we can give the correlation between the intercepts and slopes any desired value (Stoel & van den Wittenboer, 2001).

Table 5.5 illustrates this point, showing the effect of different scalings of the time variable on the coefficients of model 5. In model 5a, the time variable is scaled as in all our analyses so far, with the first measurement occasion coded as zero. In model 5b, the time variable is coded with the last measurement occasion coded as zero, and the earlier occasions with negative values −5, . . . , −1. In model 5c, the time variable is centered on its overall mean.

From the correlations between the intercepts and slopes for the time variable, we conclude in model 5b that students who end with a relatively high GPA, on average have a steeper GPA increase over time. In the centered model, 5c, this correlation is lower, but still quite clear. If we inspect the first model 5a, which codes the first occasion as zero, we see a negative correlation, meaning that subjects with a relatively low initial GPA have a steeper growth rate. It is clear that we cannot interpret the correlation between the intercept and slopes directly. This correlation can only be interpreted in combination with the scale on which the occasion variable is defined.[2]

Note that the three models in Table 5.5 have exactly identical estimates for all parameters that do not involve the measurement occasion, and exactly the same deviance and fit measures. The models are in fact equivalent. The different ways that the time variable is coded lead to what statisticians call a *re-parameterization* of the model. The three models all describe the data equally well, and are equally valid. Nevertheless, they are not identical. The situation is comparable to viewing a landscape from different angles. The landscape does not change, but some views are more interesting than others. The important lesson here is that in repeated measures analysis, careful thought must be given to the coding of the time variable. As stated, by a judicious choice of scale, we can give the correlation between the intercept and slope residuals any value that we want. If the zero point is far outside the observed values, for instance if we code the occasions as 2004, 2005, 2006, 2007, 2008, and 2009, which does make some

[2] The point on the $t$ scale where the correlation flips from negative to positive is $t^* = t_0 - (u_{01}/u_{11})$, where $t_0$ is the current zero point on the time axis. This is also the point where the intercept variance is the lowest (Mehta & West, 2000).

*Table 5.5* Results for model 5 for different scalings of measurement occasion

| Model | M5a: first occasion = 0 | M5b: last occasion = 0 | M5c: occasions centered |
|---|---|---|---|
| **Fixed part** | Coeff. (s.e.) | Coeff. (s.e.) | |
| Intercept | 2.56 (.10) | 3.07 (.09) | 2.82 (.09) |
| Occasion | 0.10 (.006) | 0.10 (.006) | 0.10 (.006) |
| Job status | –0.13 (.02) | –0.13 (.02) | –0.13 (.02) |
| GPA highschl | 0.09 (.03) | 0.09 (.03) | 0.09 (.03) |
| Gender | 0.12 (.03) | 0.12 (.03) | 0.12 (.03) |
| **Random part** | | | |
| $\sigma_e^2$ | 0.042 (.002) | 0.042 (.002) | 0.042 (.002) |
| $\sigma_{u0}^2$ | 0.038 (.006) | 0.109 (.014) | 0.050 (.006) |
| $\sigma_{u1}^2$ | 0.004 (.001) | 0.004 (.001) | 0.004 (.001) |
| $\sigma_{u01}$ | –0.002 (.002) | 0.017 (.003) | 0.007 (.001) |
| $r_{u01}$ | –.21 | .82 | .51 |
| Deviance | 170.1 | 170.1 | 170.1 |
| AIC | 188.1 | 188.1 | 188.1 |
| BIC | 233.9 | 233.9 | 233.9 |

sense, we will get an extreme correlation. If we want to interpret the correlation between the intercepts and slopes, we must make sure that the zero point has a strong substantive meaning. Adding a graphical display of the slopes for different individuals may help to interpret the results.[3]

## 5.3 EXAMPLE WITH VARYING OCCASIONS

The data in the next example are a study of children's development in reading skill and antisocial behavior. The data are a sample of 405 children who were within the first 2 years of entry to elementary school. The data consist of four repeated measures of both the child's antisocial behavior and the child's reading recognition skills. In addition, at the first measurement occasion, measures were collected of emotional

[3] In large data sets this display will be confusing, and it is better to present a plot of a random or selected subsample of the individuals.

support and cognitive stimulation provided by the mother. Other variables are the child's gender and age and the mother's age at the first measurement occasion. The data were collected using face-to-face interviews of both the child and the mother at 2-year intervals between 1986 and 1992. Between 1986 and 1992 there was an appreciable amount of panel dropout: all $N = 405$ children and mothers were interviewed at measurement occasion 1, but on the three subsequent occasions the sample sizes were 374, 297, and 294. Only 221 cases were interviewed at all four occasions. This data set was compiled by Curran (1997) from a large longitudinal data set. The predominant dropout pattern in this data set is panel dropout, meaning that if a subject was not measured at some measurement occasion, that subject is also not measured at subsequent measurement occasions. However, a small number of subjects were not measured at one of the measurement occasions, but did return at subsequent occasions.

These data are a good example of data with varying measurement occasions. Although the measurement occasions are the same for all children, their ages are all different. The children's ages at the first measurement occasion vary from 6 to 8 years. The children's ages were coded in months, and there are 25 different values for this variable. Since each child is measured at most four times, these 25 values are best treated as a time-varying predictor indicating varying measurement occasions. Figure 5.5 shows the frequency of different ages at the start of the data collection.

It is clear that with 25 different ages and only four measurement occasions, using the real age in a MANOVA-type analysis is impossible, because using listwise deletion would leave no cases to analyze. Restructured in the 'long' or 'stacked' format, we have the children's age varying from 6 to 14 years, and 1325 out of a possible 1620 observations for reading skill available for the analysis. Figure 5.6 shows a scatterplot for reading skill by child age, with the best fitting nonlinear fit line (the *loess* fit function) added. The relationship is mostly linear, reading skill increasing with age, but with some deceleration of the upwards trend at the higher ages. The variance of reading skills increases with age, which indicates that the regression coefficient for age is likely to vary across subjects.

Before the analysis, the time-varying variable child age is transformed by subtracting 6, which makes the lowest starting age zero. In addition a new variable is calculated, which is the square of the new child age variable. It is important to obtain a good model for the trend over time, and therefore it is useful to evaluate adding nonlinear trends not only by the Wald significance test, but also by the deviance difference test (which is somewhat more accurate, also for regression coefficients) and fit indices. Since the deviance difference test is used to test regression coefficients, full maximum likelihood estimation must be used. Consistent with the scatterplot, model 1, the multilevel model for predicting reading skill by time-varying age and age squared looks promising. There is also a significant cubic trend (not reported in Table 5.6), but that is very small (regression coefficient 0.006 (s.e. .002)). For simplicity, this variable

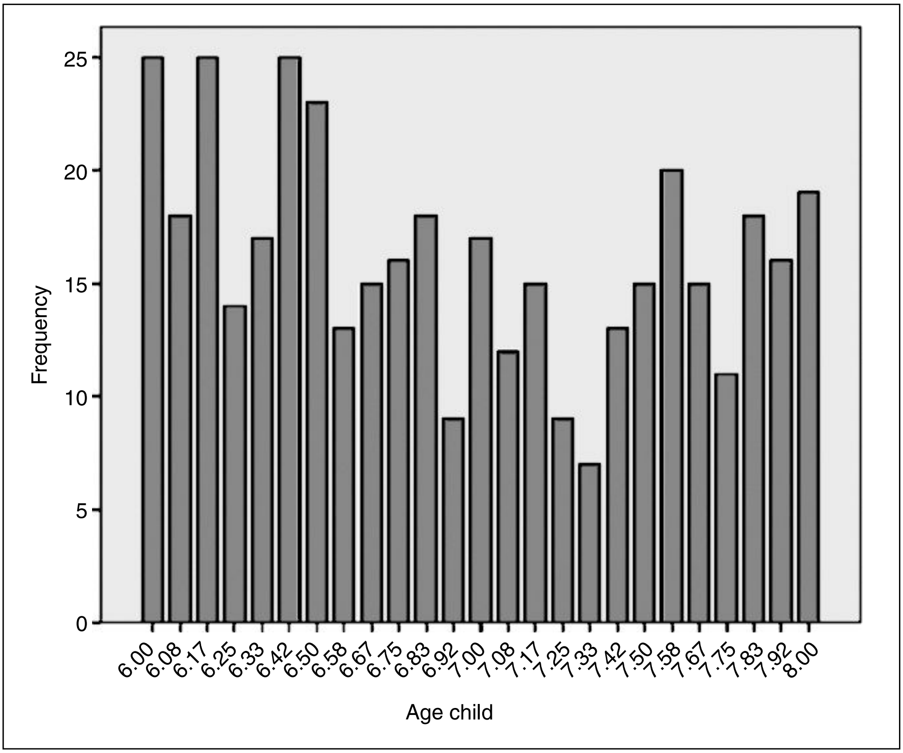


*Figure 5.5* Child ages at the first measurement occasion.

is not included in the model. Child age has a small, but significant variance across children, the squared age does not. A chi-square difference test between models 2 and 3 in Table 5.6 also indicates that the added variance and covariance are significant ($\chi^2 =$ 200.8.3, $df = 2$, $p < .001$). The correlation between the intercept and the age slope is 0.63. This indicates that children who at the initial age of 6 read comparatively well increase their reading skill faster than children who read less well at that age.

The time-invariant variables mother age, cognitive stimulation, and emotional support, which were measured at the first measurement occasion, have significant effects in the fixed model, but two of these become nonsignificant when the effect of age is assumed to vary across the children, then only cognitive stimulation has a significant regression coefficient. The deviance difference between models M2 and M3 is 200.8. As explained in Chapter 3, given that we test both an unconstrained covariance and a variance that is constrained to be non-negative, the appropriate test is a mixture of 50% chi-square with one degree of freedom and 50% chi-square with two

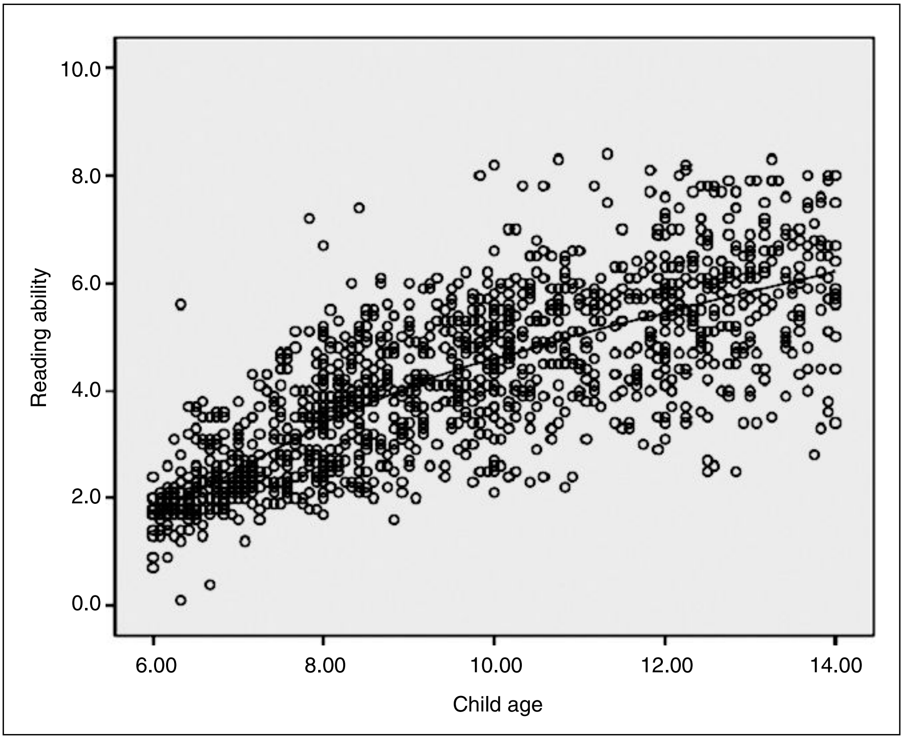


*Figure 5.6* Scatterplot of reading skill by child age.

degrees of freedom. Given the large value of $\chi^2 = 200.8$, a test at the conservative $df = 2$ produces a very small $p$-value of < .001. The decreasing AIC and BIC also suggest that model 3 is preferable. Thus, the varying effect of age belongs in the model, and the significant effect of mother age and emotional support in the fixed model is spurious.

To model the significant variance of the regression coefficient of child age, cross-level interactions of child age with the three time-invariant variables are added to the model. In this model, the interactions between the child's age and mother age and between the child's age and emotional support were significant. As a consequence, the direct effects of mother age and emotional support are retained in the model, although these direct effects are not significant by themselves. The interaction between child age and cognitive stimulation is not significant, and is dropped from the model. The last column of Table 5.6 presents the estimates for this final model. Both the deviance difference test ($\chi^2 = 20.1$, $df = 2$, $p < .001$) and the decrease in AIC and BIC indicate that model 4 is better than model 3. However, the variance of the coefficients for child

*Table 5.6* Multilevel models for reading skill

| Model | M1 | M2 | M3 | M4 |
|---|---|---|---|---|
| **Fixed part** | Coeff. (s.e.) | Coeff. (s.e.) | Coeff. (s.e.) | Coeff. (s.e.) |
| Intercept | 1.74 (.06) | 0.71 (.48) | 0.71 (.48) | 1.06 (.49) |
| Child age | 0.93 (.03) | 0.92 (.03) | 0.92 (.03) | 0.49 (.14) |
| Child age sq | –0.05 (.003) | –0.05 (.003) | –0.05 (.003) | –0.05 (.003) |
| Mother age |  | 0.05 (.02) | 0.03 (.02)$^{ns}$ | 0.02 (.02)$^{ns}$ |
| Cogn. stim. |  | 0.05 (.02) | 0.04 (.01) | 0.04 (.01) |
| Emot. support |  | 0.04 (.02) | 0.003 (.02)$^{ns}$ | –0.01 (.02)$^{ns}$ |
| Age × Momage |  |  |  | 0.01 (.005) |
| Age × Emot |  |  |  | 0.01 (.004) |
| **Random part** |  |  |  |  |
| $\sigma_e^2$ | 0.39 (.02) | 0.39 (.02) | 0.27 (.02) | 0.28 (.02) |
| $\sigma_{u0}^2$ | 0.66 (.06) | 0.60 (.04) | 0.21 (.04) | 0.21 (.04) |
| $\sigma_{u1}^2$ |  |  | 0.02 (.003) | 0.01 (.003) |
| $\sigma_{u01}$ |  |  | 0.04 (.001) | 0.04 (.001) |
| $r_{u01}$ |  |  |  | .64 |
| Deviance | 3245.0 | 3216.2 | 3015.4 | 2995.3 |
| AIC | 3255.0 | 3232.2 | 3035.4 | 3019.3 |
| BIC | 3281.0 | 3273.7 | 3087.3 | 3081.6 |

age is the same to two decimal places; when more decimals are carried it turns out that the two interaction effects explain only 0.9% of the slope variance.

The coefficients for the interactions are both 0.01. This means that when the mother age is higher, or the emotional support is high, the reading skill increases faster with the child's age. Plots are useful to interpret such interactions, Figure 5.7 shows the estimated fit lines for mothers of different ages (the range in the data is 21–29) and low vs. high emotional support (the range in the data is 0–13). Figure 5.7 illustrates that for older mothers the increase in reading skill is steeper and the leveling off less sharp. The same holds for high emotional support; with high emotional support the increase in reading skill is steeper and the leveling off less sharp. Note that the curves in Figure 5.7 are the predicted outcomes disregarding all variables not involved in the cross-level interaction. It shows the theoretical effect of the interaction, not the trend in the actual data. The apparent downward trend at the higher ages is the result of extrapolating the quadratic trend.

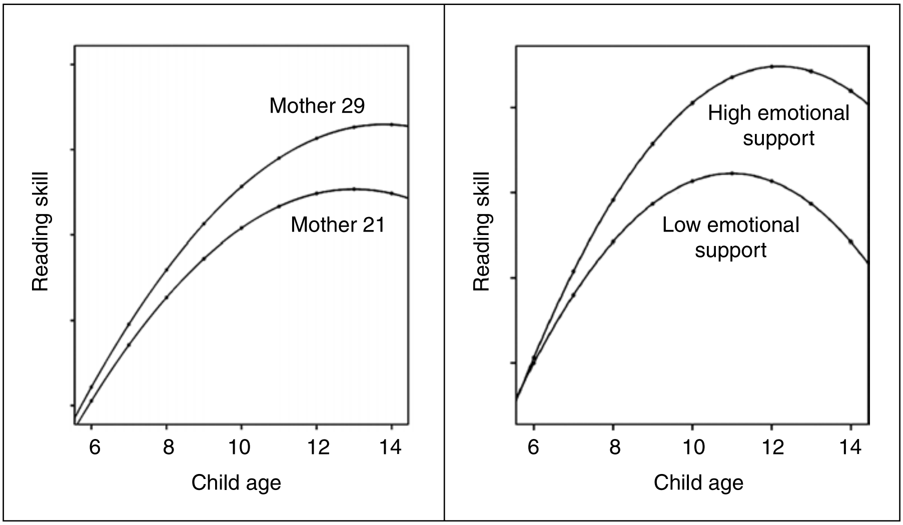


*Figure 5.7* Estimated fit lines for reading by mothers of different ages and low vs. high emotional support.

## 5.4 ADVANTAGES OF MULTILEVEL ANALYSIS FOR LONGITUDINAL DATA

Using multilevel models to analyze repeated measures data has several advantages. Bryk and Raudenbush (1992) mention five key points. First, by modeling varying regression coefficients at the measurement occasion level, we have growth curves that are different for each subject. This fits in with the way individual development is generally conceptualized (see Willett, 1988). Second, the number of repeated measures and their spacing may differ across subjects. Other analysis methods for longitudinal data cannot handle such data well. Third, the covariances between the repeated measures can be modeled as well, by specifying a specific structure for the variances and covariances at either level. This approach will be discussed in section 5.5. Fourth, if we have balanced data and use RML estimation, the usual analysis of variance based *F*-tests and *t*-tests can be derived from the multilevel regression results (see Raudenbush, 1993a). This shows that analysis of variance on repeated measures is a special case of the more general multilevel regression model. Fifth, in the multilevel model it is simple to add higher levels, to investigate the effect of family or social groups on individual development. A sixth advantage, not mentioned by Bryk and Raudenbush, is that it is straightforward to include time-varying or time-constant explanatory variables to the model, which allows us to model both the average group development and the development of different individuals over time.

## 5.5 COMPLEX COVARIANCE STRUCTURES

If multilevel modeling is used to analyze longitudinal data, the variances and covariances between different occasions have a very specific structure. In a two-level model with only a random intercept at both levels, the variance at any measurement occasion has the value $\sigma_e^2 + \sigma_{u_0}^2$, and the covariance between any two measurement occasions has the value $\sigma_{u_0}^2$. Thus, for the GPA example data, a simple linear trend model as specified by equation 5.1 is:

$$GPA_{ti} = \beta_{00} + \beta_{10}\text{Occasion}_{ti} + u_{0i} + e_{ti}, \tag{5.5}$$

where the residual variance on the occasion level is given by $\sigma_e^2$, and the residual error on the subject level is given by $\sigma_{u_0}^2$. For this and similar models without additional random effects, the matrix of variances and covariances among the occasions is given by (Goldstein, 2003; Raudenbush & Bryk, 2002):

$$\boldsymbol{\Sigma}(\mathbf{Y}) = \begin{pmatrix} \sigma_e^2 + \sigma_{u_0}^2 & \sigma_{u_0}^2 & \sigma_{u_0}^2 & \cdots & \sigma_{u_0}^2 \\ \sigma_{u_0}^2 & \sigma_e^2 + \sigma_{u_0}^2 & \sigma_{u_0}^2 & \cdots & \sigma_{u_0}^2 \\ \sigma_{u_0}^2 & \sigma_{u_0}^2 & \sigma_e^2 + \sigma_{u_0}^2 & \cdots & \sigma_{u_0}^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{u_0}^2 & \sigma_{u_0}^2 & \sigma_{u_0}^2 & \cdots & \sigma_e^2 + \sigma_{u_0}^2 \end{pmatrix}. \tag{5.6}$$

In the covariance matrix 5.6, all variances are equal, and all covariances are equal. This shows that the standard multilevel model, with a single error term at the occasion and at the subject level, is the same as assuming *compound symmetry*, the same restrictive assumption that is made in univariate analysis of variance for repeated measures. According to Stevens (2009), if the assumption of compound symmetry is violated, the standard ANOVA significance tests are too lenient, and reject the null-hypothesis more often than is warranted. Therefore, MANOVA is preferred, which estimates all variances and covariances among occasions without restrictions.

Bryk and Raudenbush (1992, p. 132) argue that uncorrelated errors may be appropriate in short time series. However, the assumption of uncorrelated errors is not essential, because the multilevel regression model can easily be extended to include an unconstrained covariance matrix at the lowest level (Goldstein, 2003). To model correlated errors, we use a multivariate response model (treated in more detail in Chapter 10 in this book) with a full set of dummy variables indicating the six consecutive measurement occasions. Thus, if we have $p$ measurement occasions, we have $p$ dummy variables, one for each occasion. The intercept term is removed from the model, so the lowest level is empty. The dummy variables are all allowed to have random slopes at the second level. Thus, for our grade point example with six occasions, we have six dummy variables $O_1, O_2, \ldots, O_6$, and the equation for a model without additional explanatory variables becomes:

$$\begin{aligned} GPA_{ti} = {} & \beta_{10}O_{1i} + \beta_{20}O_{2i} + \beta_{30}O_{3i} + \beta_{40}O_{4i} + \beta_{50}O_{5i} + \beta_{60}O_{6i} + \\ & u_{10i}O_{1i} + u_{20i}O_{2i} + u_{30i}O_{3i} + u_{40i}O_{4i} + u_{50i}O_{5i} + u_{60i}O_{6i}. \end{aligned} \tag{5.7}$$

Having six random slopes at level 2 provides us with a 6 × 6 covariance matrix for the six occasions. This is often denoted as an unstructured model for residual errors across time; all possible variances and covariances are estimated. The unstructured model for the random part is also a saturated model; all possible parameters are estimated and it cannot fail to fit. The regression slopes $\beta_{10}$ to $\beta_{60}$ are simply the estimated means at the six occasions. Equation 5.7 defines a multilevel model that is equivalent to the MANOVA approach. Maas and Snijders (2003) discuss model 5.7 at length, and show how the familiar $F$-ratios from the MANOVA approach can be calculated from the multilevel software output. An attractive property of the multilevel approach here is that it is not affected by missing data. Delucchi and Bostrom (1999) compare the MANOVA and the multilevel approach to longitudinal data using small samples with missing data. Using simulation, they conclude that the multilevel approach is more accurate than the MANOVA approach.

The model in equation 5.7 is equivalent to a MANOVA model. Since the covariances between the occasions are estimated without restrictions, it does not assume compound symmetry. However, the fixed part is also fully saturated; it estimates the six means at the six measurement occasions. To model a linear trend over time, we must replace the fixed part of equation 5.7 with the fixed part for the linear trend in equation 5.5. This gives us the following model:

$$GPA_{ti} = \beta_{00} + \beta_{10}T_{ti} + u_{10i}O_{1i} + u_{20i}O_{2i} + u_{30i}O_{3i} + u_{40i}O_{4i} + u_{50i}O_{5i} + u_{60i}O_{6i}. \tag{5.8}$$

To specify model 5.8 in standard multilevel software we must specify an intercept term that has no second-level variance component and six dummy variables for the occasions that have no fixed coefficients. Some software has built-in facilities for modeling specific covariance structures over time. If there are no facilities for longitudinal modeling, model equation 5.8 requires that the regression coefficients for the occasion dummies are restricted to zero, while their slopes are still allowed to vary across individuals. At the same time, an intercept and a linear time trend are added, which may not vary across individuals. The covariance matrix between the residual errors for the six occasions has no restrictions. If we impose the restriction that all variances are equal, and that all covariances are equal, we have again the compound symmetry model. This shows that the simple linear trend model in 5.5 is one way to impose the compound symmetry structure on the random part of the model. Since model 5.5 is nested in model 5.7, we can use the overall chi-square test based on the deviance of the two models to test if the assumption of compound symmetry is tenable.

Models with a residual error structure over time as in model 5.6 are very complex, because they assume a saturated model for the error structure. If there are $k$ measurement occasions, the number of elements in the covariance matrix for the occasions is $k(k + 1)/2$. So, with six occasions, we have 21 elements to be estimated. If the assumption of compound symmetry is tenable, models based on this model (see equation 5.5) are preferable, because they are more compact. Their random part requires only two elements $\left(\sigma_e^2 \text{ and } \sigma_{u_0}^2\right)$ to be estimated. The advantage is not only that smaller models are more parsimonious, but they are also easier to estimate. However, the compound symmetry model is very restrictive, because it assumes that there is one single value for all correlations between measurement occasions. This assumption is in many cases not very realistic, because the error term contains all omitted sources of variation (including measurement errors), which may be correlated over time. Different assumptions about the *autocorrelation* over time lead to different assumptions for the structure of the covariance matrix across the occasions. For instance, it is reasonable to assume that occasions that are close together in time have a higher correlation than occasions that are far apart. Accordingly, the elements in covariance matrix $\sum$ should become smaller, the further away they are from the diagonal. Such a correlation structure is called a *simplex*. A more restricted version of the simplex is to assume that the autocorrelation between the occasions follows the model:

$$e_t = \rho\, e_{t-1} + \varepsilon_t, \tag{5.9}$$

where $e_t$ is the error term at occasion $t$, $\rho$ is the autocorrelation, and $\varepsilon_t$ is a residual error with variance $\sigma_\varepsilon^2$. The error structure in equation 5.9 is a first order autoregressive process. This leads to a covariance matrix of the form:

$$\boldsymbol{\Sigma}(\mathbf{Y}) = \frac{\sigma_\varepsilon^2}{(1-\rho^2)} \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{k-1} \\ \rho & 1 & \rho & \cdots & \rho^{k-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{k-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{k-1} & \rho^{k-2} & \rho^{k-3} & \cdots & 1 \end{pmatrix}. \tag{5.10}$$

The first term $\sigma_\varepsilon^2/(1-\rho^2)$ is a constant, and the autocorrelation coefficient $\rho$ is between

–1 and +1, but typically positive. It is possible to have second order autoregressive processes and other models for the error structure over time. The first order autoregressive model that produces the simplex in 5.10 estimates one variance plus an autocorrelation. This is just as parsimonious as the compound symmetry model, and it assumes constant variances but not constant covariances.

Another attractive and very general model for the covariances across time is to assume that each time lag has its own autocorrelation. So, all occasions that are separated by one measurement occasion share a specific autocorrelation, all occasions that are separated by two measurement occasions share a different autocorrelation, and so on. This leads to a banded covariance matrix for the occasions that is called a Toeplitz matrix:

$$\mathbf{\Sigma}(\mathbf{Y})\sigma_e^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{k-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & 1 \end{pmatrix}. \tag{5.11}$$

The Toeplitz model poses $k - 1$ unique autocorrelations. Typically, the autocorrelations with large lags are small, so they can be removed from the model.

It should be noted that allowing random slopes for the time trend variables (e.g., for the linear trend) also models a less restricted covariance matrix for the occasions. As a result, if the measurement occasion variable, or one of its polynomials, has a random slope, it is not possible to add a completely saturated MANOVA model for the covariances across measurement occasions, as in equations 5.6 and 5.8. In fact, if we have $k$ occasions, and use $k$ polynomials with random slopes, we simply have used an alternative way to specify the saturated MANOVA model of equation 5.6.

The implication is that the restrictive assumption of compound symmetry, which is implied in the straightforward multilevel analysis of repeated measures, is also diminished when random components are allowed for the trends over time. For instance, in a model with a randomly varying linear measurement occasion variable, the variance of any specific occasion at measurement occasion $t$ is given by:

$$\text{var}(Y_t) = \sigma_{u_0}^2 + \sigma_{u_{01}}(t - t_0) + \sigma_{u_1}^2(t - t_0) + \sigma_e^2, \tag{5.12}$$

and the covariance between any two specific occasions at measurement occasions $t$ and $s$ is given by:

$$\text{cov}(Y_t,\ Y_s) = \sigma_{u_0}^2 + \sigma_{u_{01}}[(t - t_0) + (s - s_0)] + \sigma_{u_1}^2(t - t_0)(s - s_0), \tag{5.13}$$

where $s_0$ and $t_0$ are the values on which the measurement occasions $t$ and $s$ are centered (if the measurement occasion variable is already centered, $t_0$ and $s_0$ may be omitted from the equation). Such models usually do not produce the simple structure of a simplex or other autoregressive model, but their random part can be more easily interpreted in terms of variations in developmental curves or growth trajectories. In contrast, complex random structures such as the autoregression or the Toeplitz are usually interpreted in terms of underlying but unknown disturbances.

The important point is that, in longitudinal data, there are many interesting models between the extremes of the very restricted compound symmetry model and the saturated MANOVA model. In general, if there are $k$ measurement occasions, any model that estimates fewer than $k(k + 1)/2$ (co)variances for the occasions represents a restriction on the saturated model. Thus, any such model can be tested against the saturated model using the chi-square deviance test. If the chi-square test is significant, there are correlations across occasions that are not modeled adequately. In general, if our interest is mostly in the regression coefficients in the fixed parts, the variances in the random part are not extremely important. A simulation study by Verbeke and Lesaffre (1997) shows that estimates of the fixed regression coefficients are not severely compromised when the random part is mildly misspecified.

Table 5.7 presents three different models using the GPA example data. The first model has a fixed slope for the measurement occasion. The second model has a random slope for the measurement occasion, and the third model has no random effects for the intercept or the measurement occasion, but models a saturated covariance matrix across the measurement occasions. For simplicity, the table only shows the variances at the six occasions, and not the covariances. Since the fixed part of the model remains unchanged, and the interest is only in modifications of the random part, REML estimation is used.

From a comparison of the deviances, it is clear that the saturated model fits better. The deviance difference test for the random coefficient model against the saturated model is significant ($\chi^2$ = 180.1, $df$ = 21, $p$ < .001), and the AIC and BIC are smaller. However, the random coefficient model estimates only four terms in the random part, and the saturated model estimates 21 terms. It would seem attractive to seek a more parsimonious model for the random part. We can also conclude that although the saturated model leads to slightly different estimates in the fixed part, the substantive conclusions are the same. Unless great precision is needed, we may decide to ignore the better fit of the saturated model, and present the model with the random slope for the measurement occasions instead.

*Table 5.7* Results for model 5 with different random parts

| Model | Occasion fixed, comp. symm. | Occasion random, comp. symm. | Occasion fixed, saturated |
|---|---|---|---|
| **Fixed part** | Coeff. (s.e) | Coeff. (s.e.) | Coeff. (s.e.) |
| Intercept | 2.64 (.10) | 2.56 (.09) | 2.50 (.09) |
| Occasion | 0.10 (.004) | 0.10 (.006) | 0.10 (.004) |
| Job status | –0.17 (.02) | –0.13 (.02) | –0.10 (.01) |
| High GPA | 0.08 (.03) | 0.09 (.03) | 0.08 (.03) |
| Sex | 0.15 (.03) | 0.12 (.03) | 0.12 (.03) |
| **Random part** | | | |
| $\sigma_e^2$ | 0.05 (.002) | 0.042 (.002) | |
| $\sigma_{u0}^2$ | 0.046 (.006) | 0.039 (.006) | |
| $\sigma_{u1}^2$ | | 0.004 (.001) | |
| $\sigma_{u01}$ | | –0.003 (.002) | |
| $\sigma_{O1}^2$ | | | 0.090 (.009) |
| $\sigma_{O2}^2$ | | | 0.103 (.010) |
| $\sigma_{O3}^2$ | | | 0.110 (.011) |
| $\sigma_{O4}^2$ | | | 0.108 (.011) |
| $\sigma_{O5}^2$ | | | 0.104 (.011) |
| $\sigma_{O6}^2$ | | | 0.117 (.012) |
| Deviance | 314.8 | 201.9 | 21.8 |
| AIC | 318.8 | 209.9 | 63.8 |
| BIC | 329.0 | 230.3 | 170.6 |

## 5.6 STATISTICAL ISSUES IN LONGITUDINAL ANALYSIS

### 5.6.1 Investigating and analyzing patterns of change

In the previous sections, polynomial curves were used to model the pattern of change over time. Polynomial curves are often used for estimating developmental curves. They are convenient, because they can be estimated using standard linear modeling procedures, and they are very flexible. If there are $k$ measurement occasions, these can always be fitted exactly using a polynomial of degree $k - 1$. In general, in the interest of parsimony, a polynomial of a lower degree would be preferred. Another advantage of polynomial approximation is that many inherently nonlinear functions can be approximated very well by a polynomial function. Nevertheless, modeling inherently nonlinear functions directly is sometimes preferable, because it may reflect some 'true' developmental process. For instance, Burchinal and Appelbaum (1991) consider the logistic growth curve and the exponential curve of special interest for developmental models. The logistic curve describes a developmental curve where the rate of development changes slowly in the beginning, accelerates in the middle, and slows again at the end. Burchinal and Appelbaum mention vocabulary growth in children as an example of logistic growth, '. . . where children initially acquire new words slowly, beginning at about 1 year of age, then quickly increase the rate of acquisition until later in the preschool years when this rate begins to slow down again' (Burchinal & Appelbaum, 1991, pp. 29–29). A logistic growth function is inherently nonlinear, because there is no transformation that makes it possible to model it as a linear model. It is harder to estimate than linear functions, because the solution must be found using iterative estimation methods. In multilevel modeling, this becomes even more difficult, because these iterations must be carried out nested within the normal iterations of the multilevel estimation method. Estimating the nonlinear function itself rather than a polynomial approximation is attractive from a theoretical point of view, because the estimated parameters have a direct interpretation in terms of the hypothesized growth process. An alternative is to use polynomial functions to approximate the true development function. Logistic and exponential functions can be well approximated by a cubic polynomial. However, the parameters of the polynomial model have no direct interpretation in terms of the growth process, and interpretation must be based on inspection of plots of the average or some typical individual growth curves. Burchinal and Appelbaum (1991) discuss these issues with examples from the field of child development. Since the available multilevel software does not support this kind of estimation, in practice polynomial approximations are commonly used.

A general problem with polynomial functions is that they often have very high correlations. The resulting collinearity problem may cause numerical problems in the estimation. If the occasions are evenly spaced and there are no missing data, transforming the polynomials to orthogonal polynomials offers a perfect solution. Tables for orthogonal polynomials are given in most handbooks on ANOVA procedures (e.g., Hays, 1994). Even if the data are not nicely balanced, using orthogonal polynomials usually reduces the collinearity problem. If the occasions are unevenly spaced, or we want to use continuous time measurements, it often helps to center the time measures in such a way that the zero point is well within the range of the observed data points. Appendix D in this book explains how to construct orthogonal polynomials for evenly spaced measurement occasions.

Although polynomial curves are very flexible, other ways of specifying the change over time may be preferable. Snijders and Bosker (1999) discuss the use of piecewise linear functions and spline functions, which are functions that break up the development curve in different adjacent pieces, each with its own development model. Pan and Goldstein (1998) present an example of a multilevel analysis of repeated data using spline functions. Cudeck and Klebe (2002) discuss modeling developmental processes that involve phases. Using random coefficients, it is possible to model different transition ages for different subjects.

If there are $k$ fixed occasions, and there is no hypothesis involving specific trends over time, we can model the differences between the occasions perfectly using $k - 1$ polynomial curves. However, in this case it is much more attractive to use simple dummy variables. The usual way to indicate $k$ categories with dummy variables is to specify $k - 1$ dummy variables, with an arbitrary category as the reference category. In the case of fixed occasion data, it is often preferable to remove the intercept term from the regression, so all $k$ dummy variables can be used to refer to the $k$ occasions. This is taken up in more detail in Chapter 10.

### 5.6.2 Missing data and panel dropout

An often-cited advantage of multilevel analysis of longitudinal data is the ability to handle missing data (Cnaan et al., 1997; Hedeker & Gibbons, 2006; Raudenbush & Bryk, 2002; Snijders, 1996). This includes the ability to handle models with varying measurement occasions. In a fixed occasions model, observations may be missing because at some measurement occasions respondents were not measured (occasional dropout or wave nonresponse) or subjects may cease to participate altogether (panel attrition or panel mortality) (de Leeuw, 2005). In MANOVA, the usual treatment of missing measurement occasions is to remove the case from the analysis, and analyze only the complete cases. Multilevel regression models do not assume equal numbers of observations, or fixed measurement occasions, so respondents with missing observations pose no special problems here, and all cases can remain in the analysis. This is an advantage because larger samples increase the precision of the estimates and the power of the statistical tests. However, this advantage of multilevel modeling does not extend to missing observations on the explanatory variables. If explanatory variables are missing, the usual treatment is again to remove the case completely from the analysis.

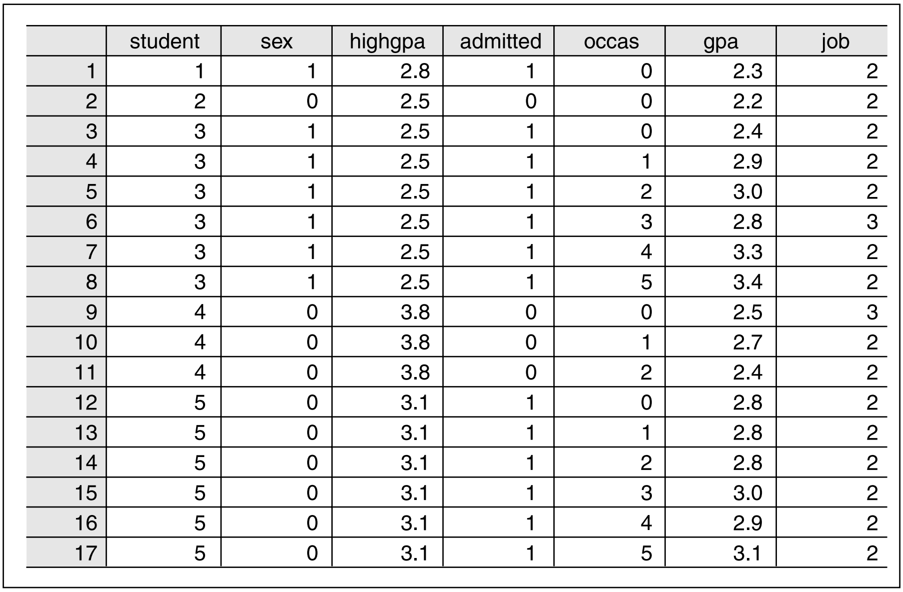
The capability to include incomplete cases in the analysis is a very important advantage. Little and Rubin (1989, 2002) distinguish between data that are missing completely at random (MCAR) and data that are missing at random (MAR). In both cases, the failure to observe a certain data point is assumed to be independent of the unobserved (missing) value. With MCAR data, the missingness must be completely independent of all other variables as well. With MAR data, the missingness may depend on other variables in the model, and through these be correlated with the unobserved values. For an accessible discussion of the differences between MAR and MCAR see Allison (2002) and McKnight, McKnight, Sidani, and Figueredo (2007).

It is clear that MCAR is a much more restrictive assumption than MAR. In longitudinal research, a major problem is the occurrence of panel attrition: individuals who after one or more measurement occasions drop out of the study altogether. Panel attrition is generally not random; some types of individuals are more prone to drop out than other individuals. In panel research, we typically have much information about the dropouts from earlier measurement occasions. In this case, it appears reasonable to assume that, conditional on these variables (which includes the score on the outcome variable on earlier occasions), the missingness is random (MAR). The complete cases method used in MANOVA assumes that data are missing completely at random (MCAR). Little (1995) shows that multilevel modeling of repeated measures with missing data assumes that the data are missing at random (MAR), provided that maximum likelihood estimation is used. Thus, MANOVA using listwise deletion leads to biased estimates when the missingness process is MAR, while multilevel analysis of data that are missing at random (MAR) leads to unbiased estimates.

Sometimes the issue arises of what to do with cases that have many missing values. For example, assume we have an experiment with an experimental group and a control group, and with a pretest before the intervention, a posttest directly after the intervention, and a follow-up test 3 months after the intervention. Some participants drop out after the pretest, so for these we have only the pretest information. Do we keep these participants in the model? The answer is *yes*. One would definitively want to include the incomplete cases in the analysis, even those with only one measurement. Deleting these is a form of listwise deletion that assumes MCAR. Keeping all incomplete cases in a multilevel analysis of these data assumes MAR. The MAR assumption is justified here because if the incomplete cases have different means on the observed variables than the complete cases, the modeling process that is based on the pattern of (co)variances (in the multilevel case also at different levels) will correct for these differences. Obviously the individuals for which there is only one measurement will provide little information, but providing that information is crucial for the justification of the MAR assumption.

An example of the bias that can be the result of analyzing MAR incomplete data with a method that assumes MCAR is presented below. In the GPA data, a substantial fraction of subjects are assigned to a panel attrition process. This attrition process is not random: if the GPA at the previous measurement occasion is comparatively low, the probability of leaving the study is comparatively high. In the resulting data set, 55% of the students have complete data, and 45% have one or more missing values for the outcome variable GPA. Figure 5.8 illustrates the structure of the data file; for subjects with missing measurement occasions the data from the available occasions are retained in the data file, and the data from the missing occasions are left out. Subsequently, these data are analyzed employing the usual multilevel analysis methods.

Table 5.8 presents the means for the six consecutive GPA measures. The first row



| | student | sex | highgpa | admitted | occas | gpa | job |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 2.8 | 1 | 0 | 2.3 | 2 |
| 2 | 2 | 0 | 2.5 | 0 | 0 | 2.2 | 2 |
| 3 | 3 | 1 | 2.5 | 1 | 0 | 2.4 | 2 |
| 4 | 3 | 1 | 2.5 | 1 | 1 | 2.9 | 2 |
| 5 | 3 | 1 | 2.5 | 1 | 2 | 3.0 | 2 |
| 6 | 3 | 1 | 2.5 | 1 | 3 | 2.8 | 3 |
| 7 | 3 | 1 | 2.5 | 1 | 4 | 3.3 | 2 |
| 8 | 3 | 1 | 2.5 | 1 | 5 | 3.4 | 2 |
| 9 | 4 | 0 | 3.8 | 0 | 0 | 2.5 | 3 |
| 10 | 4 | 0 | 3.8 | 0 | 1 | 2.7 | 2 |
| 11 | 4 | 0 | 3.8 | 0 | 2 | 2.4 | 2 |
| 12 | 5 | 0 | 3.1 | 1 | 0 | 2.8 | 2 |
| 13 | 5 | 0 | 3.1 | 1 | 1 | 2.8 | 2 |
| 14 | 5 | 0 | 3.1 | 1 | 2 | 2.8 | 2 |
| 15 | 5 | 0 | 3.1 | 1 | 3 | 3.0 | 2 |
| 16 | 5 | 0 | 3.1 | 1 | 4 | 2.9 | 2 |
| 17 | 5 | 0 | 3.1 | 1 | 5 | 3.1 | 2 |

*Figure 5.8* Example of a data set with panel attrition.

*Table 5.8* Estimated means for complete and incomplete data, six occasions

| GPA1 | GPA2 | GPA3 | GPA4 | GPA5 | GPA6 |
|---|---|---|---|---|---|
| Complete data | | | | | |
| 2.59 | 2.72 | 2.81 | 2.92 | 3.02 | 3.13 |
| Incomplete data, MANOVA (listwise $N = 109$) | | | | | |
| 2.71 | 2.89 | 2.98 | 3.09 | 3.20 | 3.31 |
| Incomplete data, multilevel model (compound symmetry) | | | | | |
| 2.59 | 2.71 | 2.81 | 2.93 | 3.07 | 3.18 |
| Incomplete data, multilevel model (saturated) | | | | | |
| 2.59 | 2.72 | 2.81 | 2.92 | 3.02 | 3.13 |

of numbers is the observed means in the complete data set. The second row is the observed means in the incomplete data set, as produced by MANOVA, using listwise deletion of incomplete cases. Compared to the complete data there is a clear upwards bias, especially in the last measurements. Using multilevel modeling results in less biased estimates when the compound symmetry model is applied to the random part, and to perfect estimates (to two decimal places) when the saturated model is applied to the random part. The difference between the outcomes of the two multilevel models emphasizes the importance of specifying a well-fitting model for the random part when there is panel attrition.

Hedeker and Gibbons (1997, 2006) present a more elaborate way to incorporate the missingness mechanism in the model. Using multilevel analysis for repeated measures, they first divide the data into groups according to their missingness pattern. Subsequently, variables that indicate these groups are included in the multilevel model as explanatory variables. The resulting *pattern mixture model* makes it possible to investigate if there is an effect of the different missing data patterns on the outcome, and to estimate an overall outcome across the different missingness patterns. This is an example of an analysis that models a specific hypothesis about data that are assumed not missing at random.

### 5.6.3 Accelerated designs

One obvious issue in longitudinal studies is that the data collection process takes a long time. Given that multilevel analysis of longitudinal data does not assume that all subjects are measured on the same occasions, it is possible to speed up the process. Different age cohorts of subjects are followed for a relatively short period of time, and then a curve is modeled across the entire age span of the data. This is called a cohort-sequential design, or an accelerated design. Figure 5.9 illustrates this design.

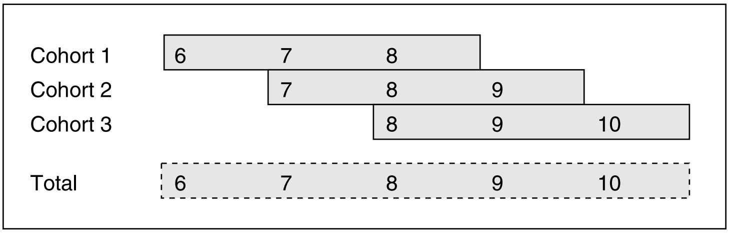


*Figure 5.9* Example of an accelerated design.

In the cohort-sequential design depicted in Figure 5.9, there are three age cohorts of children who at the beginning of the study are 6, 7, and 8 years old, respectively. The data collection takes 2 years, with data collected from the same children yearly. In the total sample, we have an age range of 6–10, and we can fit a growth curve across 5 years, although the actual data collection takes only 2 years. The reading skill data presented in the section on varying occasions is another example of accelerated design. Although the data collection takes 6 years, the age range in the sample is 6–14 years.

In an accelerated design, the growth curve is estimated on a combination of cross-sectional and longitudinal information. Obviously, this assumes that the different cohorts are comparable, for example that in Figure 5.9 the 8-year-olds in cohort 1 are comparable to the 8-year-olds in cohort 3, who are in fact 2 years older than the children in cohort 1 and are measured on a different measurement occasion. If the data collection contains a sufficient number of measurement occasions, this assumption can be tested, for example by fitting three separate linear growth curves and testing if these are equal in the three cohorts. Duncan et al. (2006) provide a discussion of cohort-sequential designs and their analysis in the context of latent curve modeling, Raudenbush and Chan (1993) discuss the analysis of cohort-sequential designs using multilevel regression models more closely, and Miyazaki and Raudenbush (2000) discuss tests for age × cohort interactions in accelerated designs.

### 5.6.4 The metric of time

In section 5.2 the measurement occasions are numbered 0, . . . , 5 to ensure that the intercept represents the starting point of the data collection. In section 5.3 on varying occasions, there are four measurement occasions, but instead of 0, . . . , 3 the real ages of the children on these occasions are used, transformed so that the youngest recorded age (6 years) equals zero. This difference points toward a larger issue: What is the correct metric of time? Developmental models, for example growth curve models, are modeling a growth process that occurs in real time. The goal of such models is to estimate the overall growth pattern, and individual deviations from that pattern in the form of individual growth curves. Counting and indexing measurement occasions does not address these goals, but using age-based models does. As discussed in the previous section on accelerated designs, age-based modeling is not unproblematic if subjects differ in age at the beginning of the data collection. On the one hand, this offers the means to analyze a wider age range than the number of years the data collection lasts, but on the other hand we must assume that there are no cohort effects. Especially when the initial age range is large, and the data collection period is short with respect to this age range, cohort effects can lead to misleading results. If there are cohort effects, it may be better to analyze the data ordered by data collection occasion, with differences in age at the first measurement occasion as predictors. A careful check of the assumption that there are no cohort effects is important in such cases.

Setting the first measurement occasion to zero, or using a transformed age measure as the measure of time, is not necessarily optimal. In the reading skill example in section 5.3, the age is transformed by subtracting six. This is reasonable; most children have some reading ability at that age. Using raw age is not reasonable, because in this metric the intercept represents reading skill at age zero, and the second-level variance represents the variation in reading skill at age zero. These estimates are clearly meaningless. A similar example is the multilevel model for the development of speech in young children in Raudenbush and Bryk (2002). Here the age is given in months, and Raudenbush and Bryk subtract 12 because the average child starts to use words at about 12 months age. Deciding on the metric of time is not only a statistical problem; it is strongly connected to the topic of study. The major substantive problem is to establish a metric of time that fits the process. For example, in a study on the importance of a major life event, such as marriage or divorce, or the birth or death of a relative, the time of this event is often set to zero. This assumes that the event can be considered to be the start of the process that is examined. If the initial point is an event, age is often added as a subject-level variable. This assumes that the curve after the event may differ for different age cohorts. Again, deciding on the metric of time is a substantive and theoretical issue, which cannot be settled by statistical reasoning alone.

## 5.7 SOFTWARE ISSUES

The models that include complex covariance structures require multilevel software that allows restrictions on the random and fixed part. Some programs (HLM, SuperMix, Prelis, and the general packages SAS, SPSS and STATA) recognize the structure of longitudinal data, and allow direct specification of various types of autocorrelation structures. For a discussion of some of these structures in the context of multilevel longitudinal models see Hedeker and Gibbons (2006). If there are many different and differently spaced occasions, MANOVA and related models become impractical. With varying occasions, it is still possible to specify an autocorrelation structure, but it is more difficult to interpret than with fixed occasions. The program MLwiN can model very general autocorrelation. Examples of such analyses are given by Goldstein, Healy, and Rasbash (1994) and Barbosa and Goldstein (2000). It should be noted that many of these programs, when the time structure is generated automatically, number the occasions starting at one. Since this makes 'zero' a non-existent value for the measurement occasion variable, this is an unfortunate choice. If this happens, software users should override it with a choice that makes better sense given their substantive question.

# 6

# The Multilevel Generalized Linear Model for Dichotomous Data and Proportions

The models discussed so far assume a continuous dependent variable and a normal error distribution. If the dependent variable is a scale in which the responses to a large number of questions are summated to one score, the data generally approximate normality. However, there are situations in which the assumption of normality is clearly violated. For instance, in cases where the dependent variable is a single dichotomous variable, both the assumption of continuous scores and the normality assumption are obviously violated. If the dependent variable is a proportion, the problems are less severe, but the assumptions of continuous scores and normality are still violated. Also, in both cases, the assumption of homoscedastic errors is violated. This chapter treats multilevel models for these kinds of data.

## 6.1 GENERALIZED LINEAR MODELS

The classical approach to the problem of non-normally distributed variables and heteroscedastic errors is to apply a transformation to achieve normality and reduce the heteroscedasticity, followed by a straightforward analysis with ANOVA or multiple regression. To distinguish this approach from the generalized linear modeling approach explained later in this chapter, where the transformation is part of the statistical model, it is often referred to as an empirical transformation. Some general guidelines for choosing a suitable transformation have been suggested for situations in which a specific transformation is often successful (e.g., Kirk, 1982; Mosteller & Tukey, 1977). For instance, for the proportion $p$ some recommended transformations are: the arcsine transformation $f(p) = 2 \arcsin(\sqrt{p})$, the logit transformation $f(p) = \text{logit}(p) = \ln(p/(1 - p))$, where 'ln' is the natural logarithm, and the probit or inverse normal transformation $f(p) = \Phi^{-1}(p)$, where $\Phi^{-1}$ is the inverse of the standard normal distribution. Thus, for proportions, we can use the logit transformation, and use standard regression procedures on the transformed variable:

$$\text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e \qquad (6.1)$$

When the dependent variable is a frequency count of events with a small probability, such as the number of errors made in a school essay, the data tend to follow a Poisson distribution, which can often be normalized by taking the square root of the scores: $f(x) = \sqrt{x}$. When the data are highly skewed, which is usually the case if, for instance, reaction time is the dependent variable, a logarithmic transformation is often used: $f(x) = \ln(x)$, or the reciprocal transformation: $f(x) = 1/x$. For reaction times the reciprocal transformation has the nice property that it transforms a variable with an obvious interpretation (reaction time) into another variable with an equally obvious interpretation (reaction speed).

Empirical transformations have the disadvantage that they are ad hoc, and may encounter problems in specific situations. For instance, if we model dichotomous data, which are simply the observed proportions in a sample of size 1, both the logistic and the probit transformations break down, because these functions are not defined for values 0 and 1. In fact, *no* empirical transformation can ever transform a dichotomous variable, which takes on only two values, into any resemblance of a normal distribution.

The modern approach to the problem of non-normally distributed variables is to include the necessary transformation and the choice of the appropriate error distribution (not necessarily a normal distribution) explicitly in the statistical model. This class of statistical models is called *generalized linear models* (Gill, 2000; McCullagh & Nelder, 1989). Generalized linear models are defined by three components:

1. an outcome variable $y$ with a specific error distribution that has mean $\mu$ and variance $\sigma^2$,
2. a linear additive regression equation that produces an unobserved (latent) predictor $\eta$ of the outcome variable $y$,
3. a *link function* that links the expected values of the outcome variable $y$ to the predicted values for $\eta$: $\eta = f(\mu)$ (McCullagh & Nelder, 1989, p. 27).

If the link function is the identity function ($f(x) = x$) and the error distribution is normal, the generalized linear model simplifies to standard multiple regression analysis. This means that the familiar multiple regression model can be expressed as a special case of the generalized linear model by stating that:

1. the probability distribution is normal with mean $\mu$ and variance $\sigma^2$, usually formulated as $y \sim \mathrm{N}(\mu,\sigma^2)$,
2. the linear predictor is the multiple regression equation for $\eta$, e.g., $\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2$,
3. the link function is the identity function given by $\eta = \mu$.

The generalized linear model separates the error distribution from the link function.

As a result, generalized linear models make it possible to extend standard regression models in two different ways: by choosing a non-normal error distribution and by using nonlinear link functions. This is nearly the same as carrying out an empirical transformation on the response variable. However, if we carry out a standard regression analysis after transforming the outcome variable, we automatically assume that the error distribution is normal on the transformed scale. But the error distribution may not be simple, or the variance may depend on the mean. Generalized linear models can deal with such situations. For instance, a commonly used generalized linear model for dichotomous data is the logistic regression model specified by:

1. the probability distribution is binomial ($\mu$) with mean $\mu$,
2. the linear predictor is the multiple regression equation for $\eta$, e.g., $\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2$,
3. the link function is the logit function given by $\eta = \text{logit}(\mu)$.

Note that this specification does not include a term for the variance of the error distribution. In the binomial distribution, the variance is a function of the mean, and it cannot be estimated separately.

The estimation method in generalized linear models is a maximum likelihood procedure that uses the inverse of the link function to predict the response variable. The inverse function for the logit used above for binomial data is the logistic transformation given by $g(x) = e^x/(1 + e^x)$. The corresponding regression model is usually written as:

$$y = \frac{e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}}{1 + e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}}.$$

This regression equation is sometimes written as $y = \text{logistic}\ (\beta_0 + \beta_1 X_1 + \beta_2 X_2)$. This is simpler, but does not show that the outcome has a binomial distribution. In reporting the results of an analysis with a generalized linear model, it is usual to list the three components of the generalized linear model explicitly. Using the regression equation $y = \text{logistic}\ (\beta_0 + \beta_1 X_1 + \beta_2 X_2)$ for estimation makes clear why modeling dichotomous data now works. Generalized linear modeling does not attempt to apply a logit transformation to the observed values 0 and 1, which is impossible, but applies the inverse logistic transformation to the expected values, which does work.

In principle, many different error distributions can be used with any link function. Many distributions have a specific link function for which sufficient statistics exist, which is called the *canonical link* function. Table 6.1 presents some commonly used canonical link functions and the corresponding error distribution.

*Table 6.1* Some canonical link functions and corresponding error distributions

| Response | Link function | Name | Distribution |
|---|---|---|---|
| Continuous | $\eta = \mu$ | Identity | Normal |
| Proportion | $\eta = \ln(\mu/(1-\mu))$ | Logit | Binomial |
| Count | $\eta = \ln(\mu)$ | Log | Poisson |
| Positive | $\eta = \mu^{-1}$ | Inverse | Gamma |

The canonical link has some desirable statistical properties, and McCullagh and Nelder (1989, Chapter 2) express a mild preference for using canonical links. However, there is no compelling reason to confine oneself to canonical link functions. Other link functions may even be better in some circumstances. For instance, although the logit link function is an appropriate choice for proportions and dichotomous data, we have the choice to specify other functions, such as the probit or the log-log function. Usually, when a link function is used, the transformation extends over the entire real line from minus to plus infinity, so there are no constraints on the values predicted by the linear regression equation. The link function is often the inverse of the error distribution, so we have the logit link for the logistic distribution, the probit link for the normal distribution, and the complementary log-log link for the extreme value (Weibull) distribution. For a discussion of error distributions and link functions in generalized linear models I refer to McCullagh and Nelder (1989).

Figure 6.1 shows the relation between the values of the proportion $p$ and the transformed values using either a logit or a probit transformation. The top shows a plot of the standard normal and logit transformation against the proportion $p$. It shows that the logit and probit transformation have a similar shape, but the standard logit transformation results in a larger spread. The bottom shows a plot where the logit transformation is standardized to have a variance of 1. It is clear from the bottom of Figure 6.1 that the logit and the probit are extremely similar.

Compared to the probit, the logit transformation has a higher peak and heavier tails, and spreads the proportions close to .00 or 1.00 over a somewhat wider range of the transformed scale. The main point of Figure 6.1 is that the differences are extremely small. Therefore, logistic and probit regression produce results that are very similar. Logistic models are more commonly used than probit models, because the exponentiated logistic coefficients can be interpreted directly as odds ratios. Other transformations, such as the log-log transformation, which is given by $f(p) = -\log(-\log(p))$, and the complementary log-log transformation, which is given by $f(p) = \log(-\log(1 - p))$, are sometimes used as well. These functions are asymmetric. For

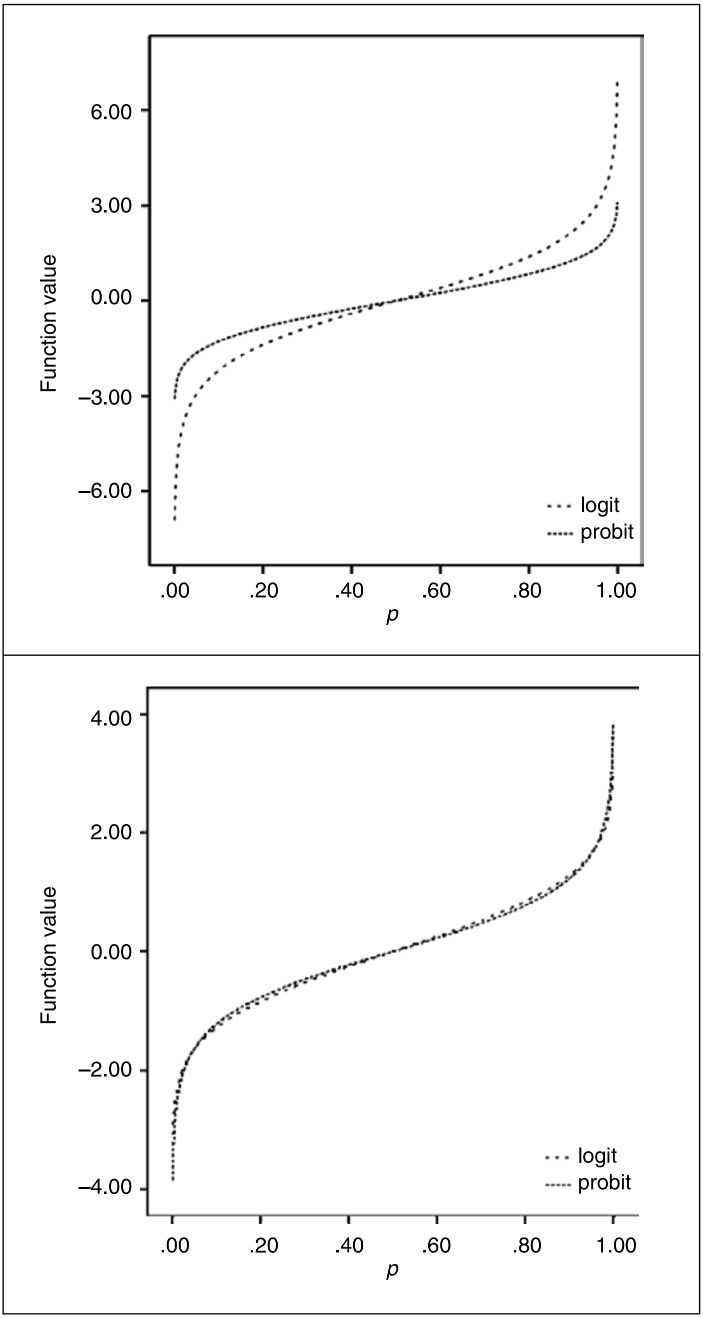


*Figure 6.1* Plot of logit and probit transformed proportions.

instance, if the proportions are larger than .5, the log-log function behaves much like the logit, while for proportions smaller than .5, it behaves more like the probit. The complementary log-log function behaves in the opposite way. McCullagh and Nelder (1989) and Aitkin, Anderson, Francis, and Hinde (1989) discuss a broad range of link functions and error distributions for various modeling problems. Again, McCullagh and Nelder (1989, pp. 108–110) express a mild preference for the canonical logit link function. Agresti (1984) discusses substantive reasons for preferring certain link functions and distributions.

Since the standard logit distribution has a standard deviation of $\sqrt{\pi^2/3} \approx 1.8$, and the standard normal distribution has a standard deviation of 1, the regression slopes in a probit model are generally close to the regression slopes in the corresponding logit model divided by 1.6–1.8 (Gelman & Hill, 2007). As the standard errors are on the same probit or logit scale, the $p$-values for the significance tests of the probit and logit regression slopes are also very similar. In many cases, the choice for a specific transformation is in practice not important. When the modeled proportions are all between .1 and .9, the differences between the logit and the probit link functions are negligible. One would need many observations with proportions close to zero or one to detect a difference between these models.

## 6.2 MULTILEVEL GENERALIZED LINEAR MODELS

Wong and Mason (1985), Gibbons and Bock (1987), Longford (1993), Goldstein (1991, 2003), and Raudenbush and Bryk (2002) describe the multilevel extension of generalized linear models. In multilevel generalized linear models, the multilevel structure appears in the linear regression equation of the generalized linear model. Thus, a two-level model for proportions is written as follows (see equation 2.5):

1. the probability distribution for $\pi_{ij}$ is binomial $(\mu, n_{ij})$ with mean $\mu$,
2. the linear predictor is the multilevel regression equation for $\eta$, e.g., $\eta = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + \gamma_{11}X_{ij}Z_j + u_{1j}X_{ij} + u_{0j}$,
3. the link function is the logit function given by $\eta = \text{logit}(\mu)$.

These equations state that our outcome variable is a proportion $\pi_{ij}$, that we use a logit link function, and that conditional on the predictor variables we assume that $\pi_{ij}$ has a binomial error distribution, with expected value $\mu$ and number of trials $n_{ij}$. If there is only one trial (all $n_{ij}$ are equal to 1), the only possible outcomes are 0 and 1, and we are modeling dichotomous data. This specific case of the binomial distribution is called the *Bernoulli* distribution. Note that the usual lowest-level residual variance $e_{ij}$ is not in the model equation, because it is part of the specification of the error distribution. If the error distribution is binomial, the variance is a function of the population proportion $\pi_{ij}$: $\sigma^2 = (\pi_{ij}/(1 - \pi_{ij})$ and it does not have to be estimated separately. Some software allows the estimation of a scale factor for the lowest-level variance. If the scale factor is set to 1, the assumption is made that the observed errors follow the theoretical binomial error distribution exactly. If the scale factor is significantly higher or lower than 1, there is *overdispersion* or *underdispersion*. Under- and overdispersion can only be estimated if the number of trials is larger than 1; in a Bernoulli distribution overdispersion cannot be estimated (McCullagh & Nelder, 1989, p. 125; Skrondal & Rabe-Hesketh, 2004, p. 127). The presence of underdispersion often indicates a misspecification of the model, such as the omission of large interaction effects. Overdispersion can occur if there are extreme outliers, or if we omit important random effects or even an entire level in a multilevel model. Very small group sizes (around three or less) also lead to overdispersion (Wright, 1997).

When overdispersion or underdispersion is present, standard errors need to be adjusted (Gelman & Hill, 2007, p. 115). The inclusion of a scale factor for under- or overdispersion improves the model fit, and corrects the standard errors. Although this takes care of the problem, it does not identify the cause of the misfit. If the scale factor is very different from 1, it is good practice to examine the problem, and to attempt to deal with it in a more explicit manner by modifying the model.

### 6.2.1 Estimation in generalized multilevel models

The parameters of generalized linear models are estimated using maximum likelihood methods. Multilevel models are also generally estimated using maximum likelihood methods, and combining multilevel and generalized linear models leads to complex models and estimation procedures. The prevailing approach, implemented for example in MLwiN, HLM, and Prelis, is to approximate the nonlinear link by a nearly linear function, and to embed the multilevel estimation for that function in the generalized linear model. This approach is a quasi-likelihood approach, and it confronts us with two choices that must be made. The nonlinear function is linearized using an approximation known as Taylor series expansion. Taylor series expansion approximates a nonlinear function by an infinite series of terms. Often only the first term of the series is used, which is referred to as a first order Taylor approximation. When the second term is also used, we have a second order Taylor approximation, which is generally more accurate. So the first choice is whether to use a first order or a second order approximation. The second choice also involves the Taylor series expansion. Taylor series linearization of a nonlinear function depends on the values of its parameters. And this presents us with the second choice: the Taylor series expansion can use the current estimated values of the fixed part only, which is referred to as marginal quasi-likelihood (MQL), or it can be improved by using the current values of the fixed part plus the residuals, which is referred to as penalized (or predictive) quasi-likelihood (PQL).

The maximum likelihood estimation procedure in multilevel modeling proceeds iteratively, starting with approximate parameter values, which are improved in each successive iteration. Thus, the estimated parameter values change during the iterations. In consequence, the Taylor series expansion must be repeated after each run of the multilevel estimation procedure, using the current estimated values of the multilevel model parameters. This results in two sets of iterations. One set of iterations is the standard iterations carried out on the linearized outcomes, estimating the parameters (coefficients and variances) of the multilevel model. In HLM, these are called the micro iterations (Raudenbush et al., 2004). The second set of iterations uses the currently converged estimates from the micro iterations to improve the Taylor series approximation. After each update of the linearized outcomes, the micro iterations are performed again. The successive improvements of the Taylor series approximation are called the macro iterations (Raudenbush et al., 2004). When iterative methods are used, convergence is generally not guaranteed. In the quasi-likelihood approach based on Taylor series approximation, there are two sets of iterations to check for convergence problems.

Estimation procedures for generalized multilevel models that are based on Taylor series expansion are discussed by Goldstein (2003), including procedures to model extra variation at the lowest level. In simulated data sets with a dichotomous response variable, Rodriguez and Goldman (1995) show that if the groups at the lowest level are small and the random effects are large, both the fixed and the random effects are severely underestimated by the first order MQL method. Goldstein and Rasbash (1996) demonstrate that using PQL and second order estimates in such situations leads to much better estimates. Even with second order expansion and PQL, the parameter estimates are still too small. Browne (1998) has repeated their analysis, using a much larger simulation setup. The amount of bias in the Taylor expansion approach can be judged from Table 6.2, which summarizes some of Browne's findings.

From the results in Table 6.2, first order MQL estimation appears almost worthless, especially regarding the second-level variance estimate. However, Goldstein and Rasbash (1996) point out that the data structure of this specific simulation is extreme, because there are very large variances in combination with very small groups. In less extreme data sets, the bias is much smaller, and then even first order MQL produces acceptable estimates. Goldstein (2003) also mentions that using second order PQL may encounter estimation problems. This explains the choice problem. If second order estimation and penalized quasi-likelihood are always better, then why not always use these? The reason is that complex models or small data sets may pose convergence problems, and we may be forced to use first order MQL. Goldstein and Rasbash (1996)

*Table 6.2* Simulation comparing MQL and PQL (see Browne, 1998)

| True value | MQL 1 estimates | PQL 2 estimates |
|---|---|---|
| $\beta_0 = 0.65$ | 0.47 | 0.61 |
| $\beta_1 = 1.00$ | 0.74 | 0.95 |
| $\beta_2 = 1.00$ | 0.75 | 0.96 |
| $\beta_3 = 1.00$ | 0.73 | 0.94 |
| $\sigma_e^2 = 1.00$ | 0.55 | 0.89 |
| $\sigma_u^2 = 1.00$ | 0.03 | 0.57 |

suggest using bootstrap methods to improve the quasi-likelihood estimates, and Browne (1998) explores bootstrap and Bayesian methods. These approaches will be treated extensively in Chapter 13. Jang and Lim (2009) show that the biases of PQL estimates of the variance components are systematically related to the biases in the PQL estimates of the regression coefficients. They also show that the biases of the PQL variance component estimates increase as the random effects become more heterogeneous. Rodriguez and Goldman (2001) compare MQL and PQL to an exact estimation approach and bootstrapping and Bayesian methods. They conclude that PQL is a considerable improvement on MQL, and that bootstrapping or Bayesian methods can further reduce the bias. However, these methods are computationally intensive, and they recommend continued use of PQL for exploratory purposes.

It is important to note that the Taylor series approach is a quasi-likelihood method. Since the likelihood that is maximized is not the real likelihood, the test statistics based on comparing the deviances of the model (which equal minus 2 times the log likelihood) are not very accurate. The AIC and BIC indices are also based on the likelihood, and should also not be used. For testing parameter estimates when Taylor series linearization is used, the Wald test or procedures based on bootstrap or Bayesian methods are preferred.

Some software does not use the quasi-likelihood approach, but uses numerical integration of the exact likelihood function. Taylor series approximation evades this difficult computing problem by maximizing an approximate likelihood. Numerical integration maximizes the correct likelihood (Schall, 1991; Wolfinger, 1993), but uses approximate calculation methods. These calculation methods involve the numerical integration of a complex likelihood function, which becomes more complicated as the number of random effects increases. The numerical approximation becomes better when the number of quadrature points in the numerical integration are increased. Unfortunately, increasing the number of quadrature points also increases the computing time, sometimes dramatically. Most software uses by default adaptive quadrature, which means that the user-provided or default number of quadrature points are not spaced evenly, but their spacing is adapted to improve estimation.

When full maximum likelihood estimation with numerical integration is used, the test procedures and goodness of fit indices based on the deviance are appropriate. Simulation research (Diaz, 2007; Hartford & Davidian, 2000; Rodriguez & Goldman, 2001) suggests that when both approaches are feasible, the numerical integration method achieves more precise estimates. Agresti, Booth, Hobert, and Caffo (2000) also recommend using numerical integration rather than Taylor expansion when it is available.

Just like second order PQL estimation, however, the numerical integration method may encounter convergence problems with certain data (Lesaffre & Spiessens, 2001). Convergence is improved if the explanatory variables have approximately the same range, which for variables with a random slope includes zero. If numerical integration is used, it helps if the user supplies good starting values, and it is recommended to check carefully if the algorithm has indeed converged (Lesaffre & Spiessens, 2001). A good check is to increase the number of integration points from its default value (which is often fairly low). If the estimates change when the number of quadrature points are increased, the smaller number was clearly not sufficient, and a new estimation run with a still larger number of quadrature points is needed to check whether the larger number was sufficient.

## 6.3 EXAMPLE: ANALYZING DICHOTOMOUS DATA

The program HLM (Raudenbush et al., 2000) includes an example data file with a dichotomous outcome variable. These *Thailand education data* stem from a national survey of primary education in Thailand (Raudenbush & Bhumirat, 1992; see Raudenbush et al., 2004, p. 115). The outcome variable ‘repeat’ is a dichotomous variable indicating whether a pupil has repeated a grade during primary education. In this example, we use child gender (0 = female, 1 = male) and having had preschool education (0 = no, 1 = yes) as predictor variables at the child level, and school mean SES as predictor variable at the school level. As outlined in the previous section, the generalized linear model has three distinct components: (1) a specific error distribution, (2) a linear regression equation, and (3) a link function. The customary link function for binomial data is the *logit* function: logit($p$) = ln($p$/(1 − $p$)). The corresponding canonical error distribution is the binomial distribution. Following the logic of the generalized linear model, we write:

$$\text{Repeat}_{ij} = \pi_{ij}; \quad \pi \sim \text{Binomial}\ (\mu) \tag{6.2}$$

$$\pi_{ij} = \text{logistic}\ (\eta_{ij}) \tag{6.3}$$

$$\eta_{ij} = \gamma_{00} + \gamma_{10}\ Sex_{ij} + Preschool\ Educ._{ij} + MeanSes_j + u_{0j} \tag{6.4}$$

or, more concisely:

$$\pi_{ij} = \text{logistic}\ (\gamma_{00} + \gamma_{10}\ Sex_{ij} + Preschool\ Educ._{ij} + MeanSes_j + u_{0j}) \tag{6.5}$$

Equations 6.2 to 6.4 and equation 6.5 describe a generalized linear model with an outcome *repeat*, which is assumed to have a binomial distribution with mean $\mu$. Since the number of trials equals 1 in all cases, we have a dichotomous outcome variable and a Bernoulli distribution. We use a logit link function, which implies that the mean $\mu$ of this distribution is predicted using a logistic regression model. In our case, this logistic regression model includes a pupil-level variable *pupil gender* and a school-level residual variance term $u_{0j}$. The parameters of this model can be estimated using the quasi-likelihood procedure involving the Taylor series expansion approach outlined above, or using the full maximum likelihood procedure with numerical integration of the likelihood function. Table 6.3 presents the results of both approaches.

*Table 6.3* Thai educational data: Predicting repeating a grade

| Model | 1st order MQL[a] | 2nd order PQL[a] | Laplace ML[b] | Numerical[c] |
|---|---|---|---|---|
| **Predictor** | Coefficient (s.e.) | Coefficient (s.e.) | Coefficient (s.e.) | Coefficient (s.e.) |
| Intercept | −1.75 (.09) | −2.20 (.10) | −2.24 (.10) | −2.24 (.11) |
| Pupil gender | 0.45 (.07) | 0.53 (.08) | 0.53 (.07) | 0.54 (.08) |
| Preschool educ. | −0.54 (.09) | −0.63 (.10) | −0.63 (.10) | −0.63 (.10) |
| Mean SES school | −0.28 (.18) | −0.29 (.22) | −0.30 (.20) | −0.30 (.22) |
| $\sigma^2_{u0}$ | 1.16 (.12) | 1.58 (.18) | 1.28 (.22) | 1.69 (.22) |

[a] Using MLwiN; [b] using HLM; [c] using SuperMix.

As Table 6.3 shows, the different methods produce estimates that are certainly not identical. First order MQL appears to underestimate both the frequency of repeats and the effect of being a male pupil, while second order PQL estimation and the numerical estimation methods in HLM and SuperMix produce regression coefficients that are very similar. The estimates of the second-level variance are also quite different. First order MQL appears to underestimate the second-level variance, while the second order PQL and numerical integration estimates are closer. The Laplace method used in

HLM improves only the estimates for the regression coefficients; the variances are still based on Taylor series expansion. Numerical integration of the entire likelihood produces the largest estimate for the school-level variance. Given the known tendency for the quasi-likelihood approach to underestimate regression coefficients and variance components, we assume that the full maximum likelihood estimates using numerical integration, which are presented in the last column in Table 6.3, are the most accurate.

The data analyzed are dichotomous. They can also be aggregated to groups of male and female students in different schools. In that case the outcome variable is aggregated to a proportion. If the same analysis is carried out on these proportions, we get effectively the same results. Since the data file has become smaller, the analysis proceeds a little faster.

It is important to understand that the interpretation of the regression parameters reported in Table 6.3 is *not* in terms of the dichotomous outcome variable *repeat*. Instead, it is in terms of the underlying variate $\eta$ defined by the logit transformation $\eta = \text{logit}(p) = \ln (p/(1 - p))$. The predicted values for $\eta$ are on a scale that ranges from $-\infty$ to $+\infty$. The logistic function transforms these predictions into values between 0 and 1, which can be interpreted as the predicted probability that an individual pupil has repeated a class. For a quick examination of the analysis results we can simply inspect the regression parameters as calculated by the program. To understand the implications of the regression coefficients for the proportions we are modeling, it is helpful to transform the predicted logit values back to the proportion scale. For example, the results in the last column of Table 6.3 show that boys repeat grades more often than girls. But, what do the intercept of –2.24 and the regression slope of 0.54 actually mean? They predict a repeat score of –2.24 for the girls (coded zero) and (–2.24 + 0.54 =) –1.70 for the boys. This is on the underlying continuous scale. Applying the logistic transformation $g(x) = e^x/(1 + e^x)$ to these estimates produces an estimated repeat rate of 9.6% for the girls and 15.4% for the boys. These values are conditional on the other variables in the model (the predictors preschool education and school mean SES), so for these values to have meaning it is important to (grand mean) center these variables.

## 6.4 EXAMPLE: ANALYZING PROPORTIONS

The second example uses data from a meta-analysis of studies that compared face-to-face, telephone, and mail surveys on various indicators of data quality (de Leeuw, 1992; for a more thorough analysis see Hox & de Leeuw, 1994). One of these indicators is the response rate: the number of completed interviews divided by the total number of eligible sample units. Overall, the response rates differ between the three data collection methods. In addition, the response rates differ also across studies. This makes it interesting to analyze what study characteristics account for these differences.

The data of this meta-analysis have a multilevel structure. The lowest level is the 'condition level', and the higher level is the 'study level'. There are three variables at the condition level: the proportion of completed interviews in that specific condition, the number of potential respondents who are approached in that condition, and a categorical variable indicating the data collection method used. The categorical data collection variable has three categories: 'face-to-face', 'telephone', and 'mail' survey. To use it in the regression equation, it is recoded into two dummy variables: a 'telephone dummy' and a 'mail dummy'. In the mail survey condition, the mail dummy equals one, and in the other two conditions it equals zero. In the telephone survey condition, the telephone dummy equals one, and in the other two conditions it equals zero. The face-to-face survey condition is the reference category, indicated by a zero for both the telephone and the mail dummy. There are three variables at the study level: the year of publication (0 = 1947, the oldest study), the saliency of the questionnaire topic (0 = not salient, 2 = highly salient), and the way the response has been calculated. If the response is calculated by dividing the response by the total sample size, we have the completion rate. If the response rate is calculated by dividing by the sample size corrected for sampling frame errors, we have the response rate. Most studies compared only two of the three data collection methods; a few compared all three. Omitting missing values, there are 47 studies, in which a total of 105 data collection methods are compared. The data set is described in Appendix A.

The dependent variable is the response. This variable is a proportion: the number of completed interviews divided by the number of potential respondents. If we had the original data sets at the individual respondents' level, we would analyze them as dichotomous data, using full maximum likelihood analysis with numerical integration. However, the studies in the meta-analysis report the aggregated results, and we have only proportions to work with. If we were to model these proportions directly by normal regression methods, we would encounter two critical problems. The first problem is that proportions do not have a normal distribution, but a binomial distribution, which (especially with extreme proportions and/or small samples) invalidates several assumptions of the normal regression method. The second problem is that a normal regression equation might easily predict values larger than 1 or smaller than 0 for the response, which are impossible values for proportions. Using the generalized linear (regression) model for the proportion $p$ of potential respondents that are responding to a survey solves both problems, which makes it an appropriate model for these data.

The hierarchical generalized linear model for our response data can be described as follows. In each condition $i$ of study $j$ we have a number of individuals who may or may not respond. Each condition $i$ of study $j$ is viewed as a draw from a specific binomial distribution. So, for each individual $r$ in each condition $i$ of study $j$ the probability of responding is the same, and the proportion of respondents in condition $i$

of study *j* is $\pi_{ij}$. Note that we could have a model where each individual's probability of responding varies, with individual-level covariates to model this variation. Then, we would model this as a three-level model, with binary outcomes at the lowest (individual) level. Since in this meta-analysis example we do not have access to the individual data, the lowest level is the condition level, with conditions (data collection methods) nested within studies.

Let $p_{ij}$ be the observed proportion of respondents in condition *i* of study *j*. At the lowest level, we use a linear regression equation to predict logit ($\pi_{ij}$). The simplest model, corresponding to the intercept-only model in ordinary multilevel regression analysis, is given by:

$$\pi_{ij} = \text{logistic } (\beta_{0j}), \tag{6.6}$$

which is sometimes written as:

$$\text{logit } (\pi_{ij}) = \beta_{0j}. \tag{6.7}$$

Equation 6.7 is a bit misleading because it suggests that we are using an empirical logit transformation on the proportions, which is precisely what the generalized linear model avoids doing, so equation 6.6 is a better representation of our model. Note again that the usual lowest level error term $e_{ij}$ is not included in equation 6.6. In the binomial distribution the variance of the observed proportion depends only on the population proportion $\pi_{ij}$. As a consequence, in the model described by equation 6.6, the lowest-level variance is determined completely by the estimated value for $\pi_{ij}$, and therefore it does not enter the model as a separate term. In most current software, the variance of $\pi$ is modeled by:

$$\text{VAR}(\pi_{ij}) = \sigma^2 \, (\pi_{ij} \, (1 - \pi_{ij})) \, / \, n_{ij}. \tag{6.8}$$

In equation 6.8, $\sigma^2$ is not a variance, but a scale factor used to model under- or overdispersion. Choosing the binomial distribution fixes $\sigma^2$ to a default value of 1.00. This means that the binomial model is assumed to hold precisely, and the value 1.00 reported for $\sigma^2$ need not be interpreted. Given the specification of the variance in equation 6.8, we have the option to estimate the scale factor $\sigma^2$, which allows us to model under- or overdispersion.

The model in equation 6.6 can be extended with an explanatory variable $X_{ij}$ at the condition level (e.g., a variable describing the condition as a mail or as a face-to-face survey):

$$\pi_{ij} = \text{logistic } (\beta_{0j} + \beta_{1j}X_{ij}). \tag{6.9}$$

The regression coefficients $\beta$ are assumed to vary across studies, and this variation is modeled by the study-level variable $Z_j$ in the usual second-level regression equations:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\, Z_j + u_{0j} \tag{6.10}$$
$$\beta_{1\mathrm{j}} = \gamma_{10} + \gamma_{11}\, Z_j + u_{1j} \tag{6.11}$$

Substituting 6.10 and 6.11 into 6.9, we get the multilevel model:

$$\pi_{ij} = \text{logistic}\ (\gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + \gamma_{11}X_{ij}\, Z_j + u_{0j} + u_{1j}X_{ij}). \tag{6.12}$$

Again, the interpretation of the regression parameters in 6.12 is *not* in terms of the response proportions we want to analyze, but in terms of the underlying variate defined by the logit transformation logit($p$) = ln ($p$/(1 − $p$)). The logit link function transforms the proportions (between .00 and 1.00 by definition) into values on a logit scale ranging from $-\infty$ to $+\infty$. The logit link is nonlinear, and in effect it assumes that it becomes more difficult to produce a change in the outcome variable (the proportion) near the extremes of .00 and 1.00, as is illustrated in Figure 6.1. For a quick examination of the analysis results, we can simply inspect the regression parameters calculated by the program. To understand the implications of the regression coefficients for the proportions we are modeling, the predicted logit values must be transformed back to the proportion scale.

In our meta-analysis, we analyze survey response rates when available, and if these are not available, the completion rate is used. Therefore the appropriate null model for our example data is not the 'intercept-only' model, but a model with a dummy variable indicating whether the response proportion is a response rate (1) or a completion rate (0). The lowest-level regression model is therefore:

$$\pi_{ij} = \text{logistic}\ (\beta_{0j} + \beta_{1j}\, \mathit{resptype}), \tag{6.13}$$

where the random intercept coefficient $\beta_{0j}$ is modeled by:

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{6.14}$$

and the slope for the variable *resptype* by:

$$\beta_{1j} = \gamma_{10}, \tag{6.15}$$

which leads by substitution to:

$$\pi_{ij} = \text{logistic}\ (\gamma_{00} + \gamma_{10}\, \mathit{resptype} + u_{0j}). \tag{6.16}$$

Since the accurate estimation of the variance terms is an important goal in meta-analysis, the estimation method uses the restricted maximum likelihood method with second order PQL approximation. For the purpose of comparison, Table 6.4 presents for the model given by equation 6.16 the first order MQL parameter estimates in addition to the preferred second order PQL approximation. Estimates were made using MLwiN.

*Table 6.4* Null model for response rates

| | MQL 1 | PQL 2 |
|---|---|---|
| **Fixed part** | Coeff. (s.e.) | Coeff. (s.e.) |
| Intercept | 0.45 (.16) | 0.59 (.15) |
| Resptype is RR | 0.68 (.18) | 0.71 (.06) |
| **Random part** | | |
| $\sigma^2_{u0}$ | 0.67 (.14) | 0.93 (.19) |

The MQL 1 method estimates the expected response rate (RR) as (0.45 + 0.68 =) 1.13, and the PQL 2 method as (0.59 + 0.71 =) 1.30. As noted before, this refers to the underlying distribution established by the logit link function, and *not* to the proportions themselves. To determine the expected proportion, we must use the inverse transformation, the logistic function, given by $g(x) = e^x/(1 + e^x)$. Using this transformation, we find an expected response rate of 0.79 for PQL 2 estimation and 0.76 for MQL 1 estimation. This is not exactly equal to the value of 0.78 that we obtain when we calculate the mean of the response rates, weighted by sample size. However, this is as it should be, for we are using a nonlinear link function, and the value of the intercept refers to the intercept of the underlying variate. Transforming that value back to a proportion is *not* the same as computing the intercept for the proportions themselves. Nevertheless, the difference is usually rather small when the proportions are not very close to 1 or 0.

In the binomial distribution (and also in the Poisson and gamma distributions), the lowest-level variance is completely determined when the mean (which in the binomial case is the proportion) is known. Therefore, $\sigma^2_e$ has no useful interpretation in these models; it is used to define the scale of the underlying continuous variate. By default, $\sigma^2_e$ is fixed at 1.00, which is equivalent to the assumption that the binomial distribution (or other distributions such as Normal, Poisson, gamma) holds *exactly*. In some applications, the variance of the error distribution turns out to be much larger than expected; there is *overdispersion* (see Aitkin et al., 1989; McCullagh & Nelder, 1989). If we estimate the overdispersion parameter for our meta-analysis data, we find a value of 21.9 with a corresponding standard error of 2.12. This value for the scale factor is simply enormous. It indicates a gross misspecification of the model. This is indeed the case, since the model does not contain important explanatory variables, such as the data collection method. Taking into account the small samples at the condition level (on average 2.2 conditions per study), the large estimate for the extrabinomial variation is probably not very accurate (Wright, 1997). For the moment, we ignore the extrabinomial variation.

It is tempting to use the value of 1.00 as a variance estimate to calculate the intraclass correlation for the null model in Table 6.3. However, the value of 1.00 is just a scale factor. The variance of a logistic distribution with scale factor 1 is $\pi^2/3 \approx 3.29$ (with $\pi \approx 3.14$, see Evans, Hastings, & Peacock, 2000). So the intraclass correlation for the null-model is $\rho = 0.93/(0.93 + 3.29) = .22$.

The next model adds the condition-level dummy variables for the telephone and the mail condition, assuming fixed regression slopes. The equation at the lowest (condition) level is:

$$\pi_{ij} = \text{logistic } (\beta_{0j} + \beta_{1j}\, resptype_{ij} + \beta_{2j}\, tel_{ij} + \beta_{3j}\, mail_{ij}), \tag{6.17}$$

and at the study level:

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{6.18}$$
$$\beta_{1j} = \gamma_{10} \tag{6.19}$$
$$\beta_{2j} = \gamma_{20} \tag{6.20}$$
$$\beta_{3j} = \gamma_{30}. \tag{6.21}$$

Substituting equations 6.18–6.21 into 6.17 we obtain:

$$\pi_{ij} = \text{logistic } (\gamma_{00} + \gamma_{10}\, resptype_{ij} + \gamma_{20}\, tel_{ij} + \gamma_{30}\, mail_{ij} + u_{0j}). \tag{6.22}$$

Until now, the two dummy variables are treated as fixed. One could even argue that it does not make sense to model them as random, since the dummy variables are simple dichotomies that code for our three experimental conditions. The experimental conditions are under control of the investigator, and there is no reason to expect their effect to vary from one experiment to another. But some more thought leads to the conclusion that the situation is more complex. If we conduct a series of experiments, we would expect identical results only if the research subjects were all sampled from exactly the same population, and if the operations defining the experimental conditions were all carried out in exactly the same way. In the present case, both assumptions are questionable. In fact, some studies have sampled from the general population, while others sample from special populations such as college students. Similarly, although most articles give only a short description of the procedures that were actually used to implement the data collection methods, it is highly likely that they were not all identical. Even if we do not know all the details about the populations sampled and the procedures used, we may expect a lot of variation between the conditions as they were actually implemented. This would result in varying regression coefficients in our model. Thus, we analyze a model in which the slope coefficients of the dummy variables for the telephone and the mail condition are assumed to vary across studies. This model is given by:

$$\begin{aligned}
\beta_{0j} &= \gamma_{00} + u_{0j} \\
\beta_{1j} &= \gamma_{10} \\
\beta_{2j} &= \gamma_{20} + u_{2j} \\
\beta_{3j} &= \gamma_{30} + u_{3j}
\end{aligned}$$

which gives:

$$\begin{aligned}
\pi_{ij} = \text{logistic } (&\gamma_{00} + \gamma_{10}\, \mathit{resptype}_{ij} + \gamma_{20}\, \mathit{tel}_{ij} + \gamma_{30}\, \mathit{mail}_{ij} \\
&+ u_{0j} + u_{2j}\, \mathit{tel}_{ij} + u_{3j}\, \mathit{mail}_{ij}).
\end{aligned} \tag{6.23}$$

Results for the models specified by equations 6.22 and 6.23, estimated by second order PQL, are given in Table 6.5.

*Table 6.5* Models for response rates in different conditions

| | Conditions fixed | Conditions random |
|---|---|---|
| **Fixed part** | Coeff. (s.e.) | Coeff. (s.e.) |
| Intercept | 0.90 (.14) | 1.16 (.21) |
| Resptype is RR | 0.53 (.06) | 0.20 (.23) |
| Telephone | –0.16 (.02) | –0.20 (.09) |
| Mail | –0.49 (.03) | –0.57 (.15) |
| **Random part** | | |
| $\sigma^2_{u0}$ | 0.86 (.18) | 0.87 (.19) |
| $\sigma^2_{u(tel)}$ | | 0.26 (.07) |
| $\sigma^2_{u(mail)}$ | | 0.57 (.19) |

The intercept represents the condition in which all explanatory variables are zero. When the telephone dummy and the mail dummy are both zero, we have the face-to-face condition. Thus, the values for the intercept in Table 6.5 estimate the expected completion rate (CR) in the face-to-face condition, 0.90 in the fixed model. The variable *resptype* indicates whether the response is a completion rate (*resptype* = 0) or a response rate (*resptype* = 1). The intercept plus the slope for *resptype* equals 1.36 in the random slopes model. These values on the logit scale translate to an expected completion rate of 0.76 and an expected response rate of 0.80 for the average face-to-face survey. The negative values of the slope coefficients for the telephone and mail dummy variables indicate that the expected response is lower in these conditions. To find out how much lower, we must use the regression equation to predict the response in the three conditions, and transform these values (which refer to the underlying logit variate) back to proportions. For the telephone conditions, we expect a response rate of 1.16, and for the mail condition 0.79. These values on the logit scale translate to an expected response rate of 0.76 for the telephone survey and 0.69 for the mail survey.

Since we use a quasi-likelihood approach, the deviance difference test is not available for testing the significance of the study-level variances. However, using the Wald test based on the standard errors, the variances of the intercept and the conditions are obviously significant, and we may attempt to explain these using the known differences between the studies. In the example data, we have two study-level explanatory variables: year of publication, and the salience of the questionnaire topic. Since not all studies compare all three data collection methods, it is quite possible that study-level variables also explain between-condition variance. For instance, if older studies tend to have a higher response rate, and the telephone method is included only in the more recent studies (telephone interviewing is, after all, a relatively new method), the telephone condition may seem characterized by low response rates. After correcting for the year of publication, in that case the telephone response rates should look better. We cannot inspect the condition-level variance to investigate whether the higher-level variables explain condition-level variability. In the logistic regression model used here, without overdispersion, the lowest-level (condition-level) variance term is automatically constrained to $\pi^2/3 \approx 3.29$ in each model, and it remains the same in all analyses. Therefore, it is also not reported in the tables.

Both study-level variables make a significant contribution to the regression equation, but only the year of publication interacts with the two conditions. Thus, the final model for these data is given by:

$$\pi_{ij} = \text{logistic}\ (\beta_{0j} + \beta_{1j}\ resptype_{ij} + \beta_{2j}\ tel_{ij} + \beta_{3j}\ mail_{ij})$$

at the lowest (condition) level, and at the study level:

$$
\begin{aligned}
&\beta_{0i} = \gamma_{00} + \gamma_{01}\, year_j + \gamma_{02}\, saliency_j + u_{0j}\\
&\beta_{1j} = \gamma_{10}\\
&\beta_{2j} = \gamma_{20} + \gamma_{21}\, year_j + u_{2j}\\
&\beta_{3j} = \gamma_{30} + \gamma_{31}\, year_j + u_{3j},
\end{aligned}
$$

which produces the combined equation:

$$
\begin{aligned}
\pi_{ij} = \text{logistic}\ (&\gamma_{00} + \gamma_{10}\, resptype_{ij} + \gamma_{20}\, tel_{ij} + \gamma_{30}\, mail_{ij} + \gamma_{01}\, year_j + \gamma_{02}\, saliency_j\\
&+ \gamma_{21}\, tel_{ij}\, year_j + \gamma_{31}\, mail_{ij}\, year_j + u_{0j} + u_{2j}\, tel_{ij} + u_{3j}\, mail_{ij}). \qquad (6.24)
\end{aligned}
$$

Results for the model specified by equation 6.24 are given in Table 6.6. Since interactions are involved, the explanatory variable year has been centered on its overall mean value of 29.74.

*Table 6.6* Models for response rates in different conditions, with random slopes and cross-level interactions

| | No interactions | With interactions |
|---|---|---|
| **Fixed part** | Coeff. (s.e.) | Coeff. (s.e.) |
| Intercept | 0.33 (.25) | 0.36 (.25) |
| Resptype | 0.32 (.20) | 0.28 (.20) |
| Telephone | –0.17 (.09) | –0.21 (.09) |
| Mail | –0.58 (.14) | –0.54 (.13) |
| Year | –0.02 (.01) | –0.03 (.01) |
| Saliency | 0.69 (.17) | 0.69 (.16) |
| Tel × Year | | 0.02 (.01) |
| Mail × Year | | 0.03 (.01) |
| **Random part** | | |
| $\sigma^2_{u0}$ | 0.57 (.13) | 0.57 (.14) |
| $\sigma^2_{u2}$ | 0.25 (.07) | 0.22 (.07) |
| $\sigma^2_{u3}$ | 0.53 (.17) | 0.39 (.15) |

Compared to the earlier results, the regression coefficients are about the same in the model without the interactions. The variances for the telephone and mail slopes in the interaction model are lower than in the model without interactions, so the cross-level interactions explain some of the slope variation. The regression coefficients in Table 6.6 must be interpreted in terms of the underlying logit scale. Moreover, the logit transformation implies that raising the response becomes more difficult as we approach the limit of 1.00. To show what this means, the predicted response for the three methods is presented in Table 6.5 as logits (in parentheses) and proportions, both for a very salient (saliency = 2) and a non-salient (saliency = 0) questionnaire topic. To compute these numbers we must construct the regression equation implied by the last column of Table 6.6, and then use the inverse logistic transformation given earlier to transform the predicted logits back to proportions. The year 1947 was coded in the data file as zero; after centering, the year zero refers to 1977. As the expected response rates in Table 6.7 show, in 1977 the expected differences between the three data collection modes are small, while the effect of the saliency of the topic is much larger. To calculate the results in Table 6.7, the rounded values for the regression coefficients given in Table 6.6 were used.

*Table 6.7* Predicted response rates (plus logits), based cross-level interaction model. Year centered on 1977

| Topic | Face-to-face | Telephone | Mail |
|---|---|---|---|
| Not salient | 0.65 (.63) | 0.61 (.44) | 0.53 (.11) |
| Salient | 0.88 (2.01) | 0.86 (1.82) | 0.82 (1.49) |

To gain a better understanding of the development of the response rates over the years, it is useful to predict the response rates from the model and to plot these predictions over the years. This is done by filling in the regression equation implied by the final model for the three survey conditions, for the year varying from 1947 to 1998 (followed by centering on the overall mean of 29.74), with saliency fixed at the intermediate value of 1.

Figure 6.2 presents the predictions for the response rates, based on the cross-level interaction model, with the saliency variable set to intermediate. The oldest study was published in 1947, the latest in 1992. At the beginning, in 1947, the cross-level model implies that the differences between the three data collection modes were large. After that, the response rates for face-to-face and telephone surveys declined, with face-to-face interviews declining the fastest, while the response rates for mail surveys remained stable. As a result, the response rates for the three data collection modes have become more similar in recent years. Note that the response rate scale has been cut off at 65%, which exaggerates the difference in trends.

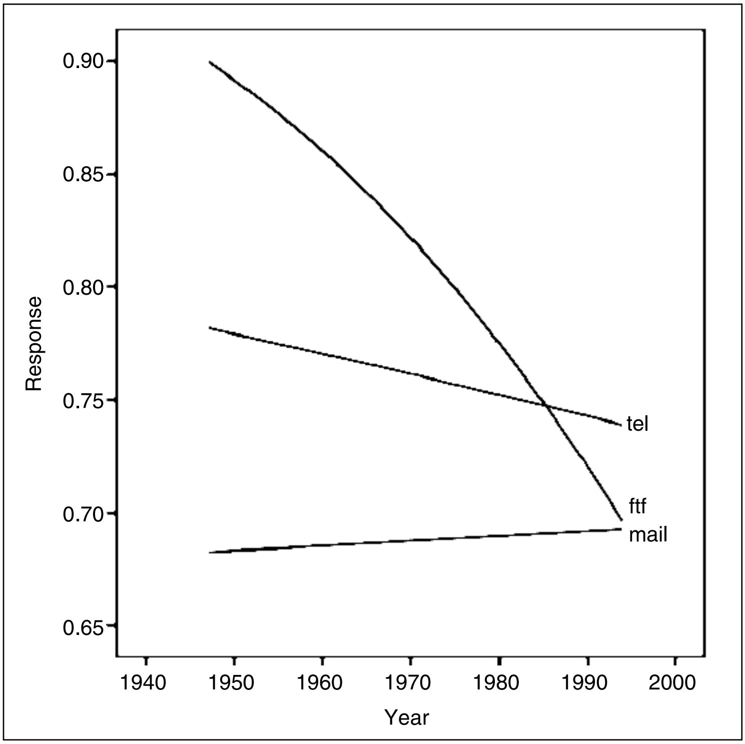


*Figure 6.2* Predicted response rates over the years.

## 6.5 THE EVER CHANGING LATENT SCALE: COMPARING COEFFICIENTS AND VARIANCES

In many generalized linear models, for example in logistic and probit regression, the scale of the unobserved latent variable $\eta$ is arbitrary, and to identify the model it needs to be standardized. Probit regression uses the standard normal distribution, with mean zero and variance one. Logistic regression uses the standard logistic distribution (scale parameter equal to one) which has a mean of zero and a variance of approximately $\pi^2/3$. The assumption of an underlying latent variable is convenient for interpretation, but not crucial.

An important issue in these models is that the underlying scale is standardized to the same standard distribution in each of the analyzed models. If we start with an intercept-only model, and then estimate a second model where we add a number of explanatory variables, we normally expect that the variance components become smaller. However, in logistic and probit regression (and many other generalized linear models), the underlying latent variable is rescaled, so the lowest-level residual variance is again $\pi^2/3$ or unity, respectively. Consequently, the values of the regression coefficients and higher-level variances are also rescaled, in addition to any real changes resulting from the changes in the model. These implicit scale changes make it impossible to compare regression coefficients across models, or to investigate how variance components change. Snijders and Bosker (1999) discuss this phenomenon briefly; a more detailed discussion is given by Fielding (2003, 2004).

The phenomenon of the change in scale is not specific to multilevel generalized linear modeling: it also occurs in single-level logistic and probit models (see Long, 1997). For the single-level logistic model, several pseudo *R*-square formulas have been proposed to provide an indication of the explained variance. These are all based on the log-likelihood. They can be applied in multilevel logistic and probit regression, provided that a good estimate of the log-likelihood is available.

A statistic that indicates the importance of each individual predictor is the partial correlation, also called Atkinson's *R*, between the outcome variable and the predictor. In logistic regression this partial correlation can be estimated using the Wald statistic and the deviance (which equals −2 times the log-likelihood). When a single predictor is assessed, the Wald test is most conveniently described as $Z_W = \beta/\text{s.e.}(\beta)$, the estimated parameter divided by its standard error. The partial correlation is then estimated as:

$$R = \sqrt{\frac{Z_w^2 - 2}{|Deviance_{\text{null}}|}}. \tag{6.25}$$

Menard (1995) warns that when the absolute value of the regression coefficient is large, the standard errors of the parameters tend to be overestimated, which makes the Wald test conservative, and in that case equation 6.25 will provide an underestimate of the partial correlation. The deviance difference test, which compares the model with and without each predictor, is considered more accurate than the Wald test. If an accurate value of the deviance is available, the $Z^2$ in equation 6.25 can be replaced by the chi-square value produced by the difference of the deviances. Since the importance of predictor variables is assessed, full maximum likelihood must be used here.

A simple analogue to the squared multiple correlation in logistic regression is McFadden's $R^2_{MF}$, given by:

$$R^2_{MF} = 1 - \frac{Deviance_{\text{model}}}{Deviance_{\text{null}}}. \tag{6.26}$$

Other pseudo *R*-squares also take the sample size into account. Common approaches to the squared multiple correlation in logistic regression are the Cox and Snell $R^2_{CS}$ and the Nagelkerke $R^2_N$. The Cox and Snell $R$-square is calculated as

$$R^2_{CS} = 1 - \exp\left(\frac{Deviance_{\text{model}} - Deviance_{\text{null}}}{n}\right), \tag{6.27}$$

where exp($x$) means $e^x$. A problem with the Cox and Snell $R^2$ is that it cannot reach the maximum value of 1. The Nagelkerke adjustment produces Nagelkerke's $R^2_N$, which can be 1:

$$R^2_N = \frac{R^2_{CS}}{1 - \exp\left(\frac{- Deviance_{\text{null}}}{n}\right)}.$$

Tabachnick and Fidell (2007) warn that these pseudo $R$-squares cannot be interpreted as explained variance, they are similar in the sense that they indicate how much of the deviance is explained, and can be used to gauge the substantive worth of the model. Although pseudo $R$-squares cannot be interpreted independently or compared across different data sets, they are useful in comparing and evaluating various models predicting the same outcome on the same data set. In general, the pseudo $R$-squares tend to be much lower than real $R$-squares, with values between .2 and .4 indicating good prediction.

The approaches discussed above can only be used if there is an accurate estimate of the likelihood. If Taylor expansion is used, the likelihood is not accurate enough for these approaches. Snijders and Bosker (1999) propose a solution for the explained variance in multilevel logistic regression that does not rely on the likelihood. It is a multilevel extension of a method proposed by McKelvey and Zavoina (1975) that is based on the explained variance of the latent outcome $\eta$ in the generalized linear model. The variance of $\eta$ is decomposed into the lowest-level residual variance $\sigma^2_R$, which is fixed to $\pi^2/3$ in the logistic and to 1 in the probit model, the second-level intercept variance $\sigma^2_{u0}$, and the variance $\sigma^2_F$ of the linear predictor from the fixed part of the model. The variance of the linear predictor is the systematic variance in the model; the other two variances are the residual errors at the two levels. The proportion of explained (modeled) variance is then given by:

$$R^2_{MZ} = \frac{\sigma^2_F}{\sigma^2_F + \sigma^2_{u0} + \sigma^2_R}. \tag{6.28}$$

The variance of the linear predictor is sometimes given by the software, but it can easily be determined by calculating the predictions from the regression equation.

Because McKelvey and Zavoina's approach is very similar to OLS *R*-squares, we can interpret the McKelvey and Zavoina $R^2$ as a multiple *R*-squared for the latent continuous variable. Long (1997) reviews several pseudo *R*-squares, and concludes that the McKelvey and Zavoina $R^2_{MZ}$ is the most attractive. The accuracy of various pseudo *R*-squares has been assessed in simulation studies by predicting a continuous variable through OLS regression and a dichotomized version by logistic regression and comparing the pseudo *R*-squares to the OLS *R*-square. In these simulations, McKelvey and Zavoina's pseudo $R^2$ was the closest to the OLS *R*-square, which the other pseudo *R*-squares tended to underestimate (see Long, 1997).

Calculating the total variance of the latent outcome is in itself useful. The square root of the total variance $\sigma^2_F + \sigma^2_{u0} + \sigma^2_R$ is after all the standard deviation of the latent variable. This can be used to compute standardized regression coefficients.

The various pseudo *R*-squares do not distinguish between different levels. However, following McKelvey and Zavoina's approach, we can calculate the total variance of the latent variable as $\sigma^2_F + \sigma^2_{u0} + \sigma^2_R$. Since rescaling takes place only when lowest-level variance is explained, only first-level predictors are used here. For the null model, the total variance is $\sigma^2_0 = \sigma^2_{u0} + \sigma^2_R$, with $\sigma^2_R \approx 3.29$. For the model *m* including the first-level predictor variables, the total variance is $\sigma^2_m = \sigma^2_F + \sigma^2_{u0} + \sigma^2_R$. Hence, we can calculate a scale correction factor that rescales the model *m* to the same underlying scale as the null model. The scale correction factor equals $\sigma_0/\sigma_m$ for the regression coefficients and $\sigma^2_0/\sigma^2_m$ for the variance components. Next, we can rescale both the regression coefficients and the variance estimates $\sigma^2_{u0}$ and $\sigma^2_R$ by the appropriate scale correction factor, which makes them comparable across models. The scale corrected variance estimates are useful for assessing the amount of variance explained separately at the different levels, using the procedures outlined in Chapter 4. Given all these prospects made possible by rescaling to the scale of the intercept-only model, the McKelvey and Zavoina method is the most appealing.

We use the Thai educational data to illustrate these procedures. The sample consists of 8582 pupils in 357 schools. The first three columns in Table 6.8 present the null model M0, the estimates of the model M1 that includes pupil gender and preschool education, and model M2 that adds school mean SES as predictor. The estimation method is full numerical integration (SuperMix with 20 adaptive quadrature points), and we report three decimal places for a more precise calculation. The fourth column in Table 6.8 shows the partial *r* for the predictors in the model M2. The largest partial *r* is .09 for pupil gender, which Cohen (1988) classifies as a small correlation. If the deviance difference is used instead of the square of the Wald *Z*, the partial correlation is estimated as .10. We can take the square root of a pseudo *R*-square and obtain the analogue to the multiple correlation. For M2 the McFadden *R* is .13, the Cox and Snell *R* is .11, and the Nagelkerke *R* is .15. All these estimates are rather

*Table 6.8* Thai educational data: Logistic regression estimates with partial *r*

| Model | M0 | M1 | M2 | M2<br>partial *r* | M2<br>standardized |
|---|---|---|---|---|---|
| **Predictor** | Coeff. (s.e.) | Coeff. (s.e.) | Coeff. (s.e.) | | |
| Intercept | −2.234 (.088) | −2.237 (.107) | −2.242 (.106) | – | – |
| Pupil gender | | 0.536 (.076) | 0.535 (.076) | .09 | .60 |
| Pre educ. | | −0.642 (.100) | −0.627 (.100) | .08 | −.69 |
| Mean SES | | | −0.296 (.217) | .01 | −.25 |
| $\sigma^2_{u0}$ | 1.726 (.213) | 1.697 (.211) | 1.686 (.210) | | |
| Deviance | 5537.444 | 5443.518 | 5441.660 | | |

similar and all lead to the same conclusion, that the explanatory power of the model on repeating a class is small.

The variance of the linear predictor for M2 (calculated in the raw data file, using the coefficients for the three predictors in the fixed part of the regression equation), $s^2_F$, is 0.201. Thus, the explained variance using the McKelvey and Zavoina method is 0.201/(0.201 + 1.686 + 3.290) = 0.039, and the corresponding multiple correlation is .20. Consistent with the existing simulation studies, it is the highest of the pseudo *R* estimates, but we still conclude that we explain only a small amount of variance. The square root of the variance of the linear predictor, which is its standard deviation, is 0.488. This value is used, together with the standard deviations of the predictors in the sample, to calculate the standardized regression coefficients for M2 in the last column of Table 6.8.

To calculate estimates of explained variance at separate levels we need to bring the models to the same scale. The total variance of the latent outcome variable in the intercept-only model is 1.726 + 3.290 = 5.016 and the corresponding standard deviation is 2.240. The variance of the linear predictor, using only the level 1 variables in model M1, is 0.171. Therefore, the total variance in the model with only the first-level predictor variables is 5.016 + 0.171 = 5.187, and the corresponding standard deviation is 2.277. The difference is tiny, and reflects the small effects of the explanatory variables on the outcome variable. The scale correction factor is 2.240/2.277 = 0.983. The variances must be scaled with the square of the scaling factor, which is 0.967. The rescaled regression coefficients are on the same scale as the intercept-only model. If we build up the model in steps, as suggested in Chapter 4, we can use this method to rescale all results to the scale of the null model, thus retaining comparability across different models. When changes are made to the fixed part at the lowest level, the scaling factor has to be calculated again. When changes are made to the higher levels, as in M2 in Table 6.7, the scaling factor remains the same, because such changes do not alter the explained variance at the lowest level. Table 6.9 presents the estimates for the models in Table 6.8 in both the raw and the rescaled version. The lowest-level variance $\sigma_R^2$ is added to the table. In the null model this has the distributional value of 3.29, but in the subsequent models it is rescaled, just like the higher-level variances.

*Table 6.9* Thai educational data: Logistic regression estimates with rescaling

| Model | M0 | M1 | SC M1 | M2 | SC M2 |
|---|---|---|---|---|---|
| **Predictor** | Coeff. (s.e.) | Coeff. (s.e.) | Coeff. (s.e.) | Coeff. (s.e.) | |
| Intercept | −2.234 (.088) | −2.237 (.107) | −2.21 (.10) | −2.242 (.106) | 2.17 (.10) |
| Pup. gender | | 0.536 (.076) | 0.52 (.07) | 0.535 (.076) | 0.52 (.07) |
| Pre educ. | | −0.642 (.100) | −0.62 (.10) | −0.627 (.100) | −0.60 (.10) |
| Mean SES | | | | −0.296 (.217) | 0.29 (.21) |
| $\sigma_R^2$ | 3.290 | n/a | 3.076 | n/a | 3.076 |
| $\sigma_{u0}^2$ | 1.726 (.213) | 1.697 (.211) | 1.587 | 1.686 (.210) | 1.577 (.20) |
| Deviance | 5537.444 | 5443.518 | – | 5441.660 | – |

In Table 6.9, the columns SC M1 and SC M2 contain the scale corrected values of the regression coefficients and variances, including the residual variance at the lowest level. Using procedures described in Chapter 4, we can estimate the explained variance in M1 and M2. In the empty model, the lowest-level variance has its distributional value 3.290. In the scale corrected M1, this variance decreases to 3.076, which leads to a level 1 explained variance in M1 of 0.065. At the second level, the variance drops from 1.726 to 1.587, which leads to a level 2 explained variance in M1 of 0.081. In the scale corrected M2, the variance goes down to 1.577, and the explained variance at the school level increases to 0.086. In terms of explained variance, the school variation in repeat rates is explained better by the differences in pupil composition, and less by the contextual effect of school mean SES. The partial correlations in the last column of Table 6.8 indicate also that schools' mean SES is relatively unimportant.

The examples in this chapter refer to logistic and probit regression, because these are the most commonly used distributions. However, almost any continuous function that maps probability onto the real line from minus infinity to plus infinity can be used to form a generalized linear model. When it is desirable to rescale such models, one needs to know the variance of the standard distribution. For example, the log-log and complementary log-log distributions that were mentioned earlier have variance $\pi^2/6$. Here the variance depends on the predicted outcome, and also changes between one model and the next. For details for a large number of distributions I refer to Evans et al. (2000).

## 6.6 INTERPRETATION AND SOFTWARE ISSUES

The models and parameter estimates described earlier are the so-called 'unit-specific' models. Unit-specific models predict the outcome of individuals and groups conditional on all random effects included in the model. The interpretation of such models is equivalent to the interpretation of effects in standard multilevel regression models; the regression coefficients for all explanatory variables reflect the predicted change in outcome when a predictor is changed by one unit. With nonlinear models, one can also estimate a population average model. This model does not condition on the random effects, but averages over all random effects. This approach is a form of estimation termed 'generalized estimating equations' (GEE), discussed in more detail in Chapter 13. In scientific research where the variation in behavior of individuals within groups is studied, and the focus is on explaining how individual and group-level variables affect the behavior, unit-specific models are appropriate. Population average models have their place in policy research, where the research problem concerns estimating the expected change in an entire population when one of the group-level variables is manipulated. For a technical discussion of the difference between unit specific and population average models I refer to Raudenbush and Bryk (2002); an accessible discussion in the context of epidemiological research is given by Hu et al. (1998).

For estimating nonlinear models, there is an increasing availability of software that can use numerical integration. The precise choices available depend on the software. As the example analyses in this chapter show, in many cases the difference between PQL and numerical integration is small, especially for the regression coefficients in the fixed part. However, this is not always the case. The simulations by Rodriguez and Goldman (1995) show that the combination of small groups and a high intraclass correlation can produce a severe bias, even when PQL is used. This combination is not unusual in some kinds of research. For instance, when families or couples are studied, the groups will be small and the intraclass correlation is often high. The smallest group sizes occur when we analyze dyadic data (couples). Given some amount of spouse nonresponse, the average group size can be well under 2 for such data, and if the model is nonlinear, PQL can be expected to produce biased results. If numerical integration is not available, bootstrap or Bayesian methods should be used to improve the estimates (see Chapter 13).

As mentioned earlier, with dichotomous data the overdispersion parameter cannot be estimated. Nevertheless, some software in fact allows the estimation of overdispersion with dichotomous data, especially when quasi-likelihood estimation methods are used. With dichotomous data, the overdispersion parameter is superfluous, and should in general not be included in the model (Skrondal & Rabe-Hesketh, 2007).

The program MLwiN does not use numerical integration, but bootstrapping and Bayesian methods are available. The multilevel software developed by Hedeker and Gibbons (see Hedeker & Gibbons, 1996a, 1996b; and Hedeker, Gibbons, du Toit, & Cheng, 2008) and the aML program developed by Lillard and Panis (Lillard & Panis, 2000) use numerical integration. The various packages by Hedeker and Gibbons are available in an integrated package called SuperMix (Hedeker et al., 2008). HLM has the option of using numerical integration for dichotomous data only (Raudenbush et al., 2004), using an approach called Laplace approximation. This is equivalent to using numerical integration with only one integration point (Rabe-Hesketh & Skrondal, 2008, p. 251). The Mplus software (Muthén & Muthén, 2007) includes several options for numerical integration. Several large software packages, such as SAS, STATA but also the freely available package **R**, have numerical integration for multilevel generalized linear models, as does the free STATA add-in GLLAMM (Rabe-Hesketh, Skrondal, & Pickles, 2004).

# 7

# The Multilevel Generalized Linear Model for Categorical and Count Data

When outcome variables are severely non-normal, the usual remedy is to try to normalize the data using a nonlinear transformation, to use robust estimation methods, or by a combination of these (see Chapter 4 for details). Then again, just like dichotomous outcomes, some types of data will always violate the normality assumption. Examples are ordered (ordinal) and unordered (nominal) categorical data, which have a uniform distribution, or counts of rare events. These outcomes can sometimes also be transformed, but they are preferably analyzed in a more principled manner, using the generalized linear model introduced in Chapter 6. This chapter describes the use of the generalized linear model for ordered categorical data and for count data.

## 7.1 ORDERED CATEGORICAL DATA

There is a long tradition, especially in the social sciences, of treating ordered categorical data as if they were continuous and measured on an interval scale. A prime example is the analysis of Likert scale data, where responses are collected on ordered response categories, for example ranging from 1 = totally disagree to 5 = totally agree. Another example is a physician's prognosis for a patient categorized as 'good', 'fair', and 'bad'.

The consequences of treating ordered categorical data as continuous are well known, both through analytical work (Olsson, 1979) and simulations (e.g., Dolan, 1994; Muthén & Kaplan, 1985). The general conclusion is that if there are at least five categories, and the observations have a symmetric distribution, the bias introduced by treating categorical data as continuous is small (Bollen & Barb, 1981; Johnson & Creech, 1983). With seven or more categories, the bias is very small. If there are four or fewer categories, or the distribution is skewed, both the parameters and their standard errors tend to have a downward bias. When this is the case, a statistical method designed for ordered data is needed. Such models are discussed by, among others, McCullagh and Nelder (1989) and Long (1997). Multilevel extensions of these models are discussed by Goldstein (2003), Raudenbush and Bryk (2002), and Hedeker and Gibbons (1994). This chapter treats the cumulative regression model, which is frequently used in practice; see Hedeker (2008) for a discussion of other multilevel models for ordered data.

### 7.1.1 Cumulative regression models for ordered data

A useful model for ordered categorical data is the cumulative ordered logit or probit model. It is common to start by assigning simple consecutive values to the ordered categories, such as 1 . . . $C$ or 0 . . . $C-1$. For example, for a response variable $Y$ with three categories such as 'never', 'sometimes', and 'always' we have three response probabilities:

$$\begin{aligned}
&\text{Prob}(Y=1)=p_1\\
&\text{Prob}(Y=2)=p_2\\
&\text{Prob}(Y=3)=p_3
\end{aligned}$$

The cumulative probabilities are given by:

$$\begin{aligned}
&p_1^*=p_1\\
&p_2^*=p_1+p_2\\
&p_3^*=p_1+p_2+p_3=1
\end{aligned}$$

where $p_3^*$ is redundant. With $C$ categories, only $C-1$ cumulative probabilities are needed. Since $p_1$ and $p_2$ are probabilities, generalized linear regression can be used to model the cumulative probabilities. As stated in Chapter 6, a generalized linear regression model consists of three components:

1. an outcome variable $y$ with a specific error distribution that has mean $\mu$ and variance $\sigma^2$,
2. a linear additive regression equation that produces a predictor $\eta$ of the outcome variable $y$,
3. a *link function* that links the expected values of the outcome variable $y$ to the predicted values for $\eta$: $\eta=f(\mu)$.

For a logistic regression we have the logit link function:

$$\eta_c=\text{logit}(p_c^*)=\ln\left(\frac{p_c^*}{1-p_c^*}\right), \tag{7.1}$$

and for probit regression the inverse normal link:

$$\eta_c = \Phi(p_c^*)^{-1}, \tag{7.2}$$

for $c = 1 \ldots C - 1$. A two-level intercept-only model for the cumulative probabilities is then written as:

$$\eta_{ijc} = \theta_c + u_{0j}. \tag{7.3}$$

Equation 7.3 specifies a different intercept $\theta_c$ for each of the estimated probabilities. These intercepts are called thresholds, because they specify the relationship between the latent variable $\eta$ and the observed categorical outcome. The position on the latent variable determines which categorical response is observed. Specifically,

$$y_i = \begin{cases} 1, & \text{if } \eta_i \leq \theta_1 \\ 2, & \text{if } \theta_1 < \eta_i \leq \theta_2 \\ 3, & \text{if } \theta_2 < \eta_i, \end{cases}$$

where $y_i$ is the observed categorical variable, $\eta_i$ is the latent continuous variable, and $\theta_1$ and $\theta_2$ are the thresholds. Note that a dichotomous variable only has one threshold, which becomes the intercept in a regression equation.

Figure 7.1 illustrates the relations between the thresholds $\theta$, the unobserved response variable $\eta$, and the observed responses. As noted by McCullagh and Nelder (1989), assuming a continuous latent distribution underlying the categorical responses is not strictly necessary for use of generalized linear regression models like the kind presented here, but it does help interpretation. Figure 7.1 shows that the standard logistic distribution has a larger variance ($\pi^2/3 \approx 3.29$) than the standard normal distribution, something that was discussed earlier in the chapter on dichotomous data. So in the logistic regression model the regression coefficients tend to be larger than in the probit model, which reflects just a difference in the scaling of the latent outcome variable. The standard errors are also larger in the logistic model, and as Figure 7.1 clearly shows, the thresholds are also scaled. Since the relative shape of the two distributions is extremely similar (see Chapter 6, Figure 6.1 for an illustration), when the results are standardized or expressed as predicted response probabilities, the results are also very similar.

The model in 7.1 assumes *parallel regression* lines; the effect of predictor variables on the regression is that the entire structure is shifted. In the logit model this translates to the proportional odds model, which assumes that the predictors have the same effect on the odds for each category $c$. The assumption of proportional odds is equivalent to the assumption of parallel regression lines; when the structure is shifted, the slope of the regression lines does not change. The same assumption is also made in the probit model.

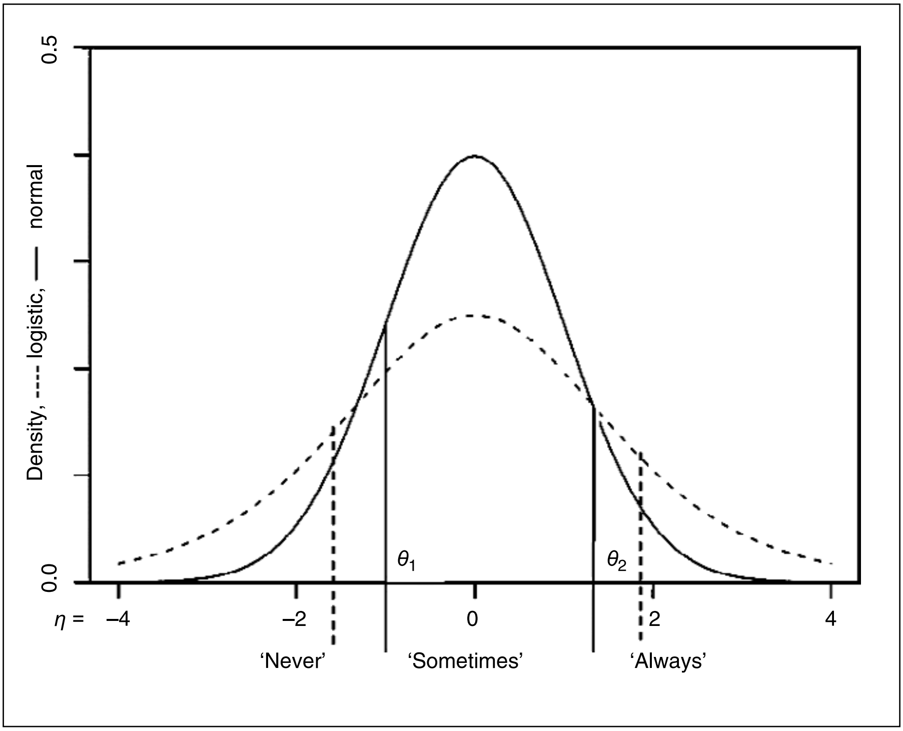


*Figure 7.1* Thresholds and observed responses for logit and probit model.

### 7.1.2 Cumulative multilevel regression models for ordered data

Just as in multilevel generalized linear models for dichotomous data, the linear regression model is constructed on the underlying logit or probit scale. Both have a mean of zero, the variance of the logistic distribution is $\pi^2/3$ (standard deviation 1.81), and the standard normal distribution for the probit has a variance of 1. As a consequence, there is no lowest-level error term $e_{ij}$, similar to its absence in generalized linear models for dichotomous data. In fact, dichotomous data can be viewed as ordered data with only two categories. Since the standard logit distribution has a standard deviation of $\sqrt{\pi^2/3} \approx 1.8$, and the standard normal distribution has a standard deviation of 1, the regression slopes in a probit model are generally close to the regression slopes in the corresponding logit model divided by 1.6–1.8 (Gelman & Hill, 2007). Assuming individuals $i$ nested in groups $j$, and distinguishing between the different cumulative proportions, we write the model for the lowest level as follows:

$$\begin{array}{l} \eta_{1ij} = \theta_{1j} + \beta_{1j} X_{ij} \\ \eta_{2ij} = \theta_{2j} + \beta_{1j} X_{ij}, \\ \vdots \quad \vdots \quad \vdots \\ \eta_{cij} = \theta_{cj} + \beta_{1j} X_{ij} \end{array} \tag{7.4}$$

where the thresholds $\theta_1 \ldots \theta_c$ are the intercepts for the response outcomes. The model given by 7.4 is problematic, because we have a collection of intercepts or thresholds that can all vary across groups. The interpretation of such variation is that the groups differ in how the values of the underlying $\eta$ variable are translated into response categories. If this is the case, there is no measurement equivalence between different groups, and it is impossible to make meaningful comparisons. Therefore the model is rewritten by subtracting the value of the first threshold from all thresholds. Thus, the first threshold becomes zero, and is thereby effectively removed from the model. It is replaced by an overall intercept, which is allowed to vary across groups. Thus, the lowest-level model becomes:

$$\begin{array}{l} \eta_{1ij} = \beta_{0j} + \beta_{1j} X_{ij} \\ \eta_{2ij} = \theta_2 + \beta_{0j} + \beta_{1j} X_{ij}, \\ \vdots \quad \vdots \quad \vdots \quad \vdots \\ \eta_{cij} = \theta_c + \beta_{0j} + \beta_{1j} X_{ij} \end{array} \tag{7.5}$$

where in 7.5 the threshold $\theta_c$ is equal to $\theta_c - \theta_1$ in 7.4. Obviously, the value for $\beta_0$ in 7.5 will be equal to $-\theta_1$ in 7.4. Note that the transformed thresholds $\theta_c$ do not have a subscript for groups; they are assumed to be fixed to maintain measurement invariance across the groups. To keep the notation simple, we will continue to use $\theta_2 \ldots \theta_c$ to refer to the thresholds in the 7.5 parameterization, where the first threshold is constrained to zero to allow an intercept in the model, and the other thresholds are all shifted.

From this point, the multilevel generalized model for ordinal observations is constructed following the accustomed procedures. Thus, the intercept $\beta_{0j}$ and the slope $\beta_{1j}$ are modeled using a set of second-level regression equations:

$$\begin{array}{l} \beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j} \\ \beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j}. \end{array} \tag{7.6}$$

The single equation version of the model is:

$$\begin{array}{l} \eta_{1ij} = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} X_{ij} Z_j + u_{0j} + u_{1j} X_{ij} \\ \eta_{2ij} = \theta_2 + \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} X_{ij} Z_j + u_{0j} + u_{1j} X_{ij}, \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ \eta_{cij} = \theta_c + \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} X_{ij} Z_j + u_{0j} + u_{1j} X_{ij} \end{array} \tag{7.7}$$

or, in a simpler notation:

$$\eta_{cij} = \theta_c + \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + \gamma_{11}X_{ij}Z_j + u_{0j} + u_{1j}X_{ij}, \tag{7.8}$$

with the condition that $\theta_1$ is zero. Using the empty model:

$$\eta_{cij} = \theta_c + \gamma_{00} + u_{0j}, \tag{7.9}$$

we obtain estimates of the variance of the residual errors $u_0$ that can be used to calculate the intraclass correlation. The residual first-level variance is equal to $\pi^2/3 \approx 3.29$ for the logit and 1 for the probit scale. Note that the ICC is defined on the underlying scale, and not on the observed categorical response scale. Just as in the dichotomous case, the underlying scale is rescaled in each model, and the regression coefficients from different models can not be compared directly.

Modeling the cumulative probabilities $p_1, p_1 + p_2, \ldots, p_1 + p_2 + \ldots + p_{c-1}$ makes the last response category the reference category. As a result, the regression coefficients in the cumulative regression model will have a sign that is the opposite of the sign given by an ordinary linear regression. This is confusing, and most model and software writers solve this effectively by writing the regression equation, for example for 7.8, as:

$$\eta_{cij} = -1(\theta_c + \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + \gamma_{11}X_{ij}Z_j + u_{0j} + u_{1j}X_{ij}), \tag{7.10}$$

which restores the signs to the direction they would have in a standard linear regression. However, this is not universally done, and software users should understand what their particular software does.

The estimation issues discussed in modeling dichotomous outcomes and proportions also apply to estimating ordered categorical models. One approach is to use Taylor series linearization, using either the marginal quasi likelihood (MQL) or the penalized quasi likelihood (PQL). PQL is generally considered more accurate, but in either case the approximation to the likelihood is not accurate enough to permit deviance difference tests. The other approach is to use numerical methods, which is more computer-intensive and more vulnerable to failure. When numerical methods are used, their performance can be improved by paying careful attention to the explanatory variables. Outliers and the use of variables that differ widely in scale increase the risk of failure. In addition, centering explanatory variables with random slopes to make sure that zero is an interpretable value is important. The next section presents an example where these issues are present and are dealt wth.

It should be noted that the proportional odds assumption is often violated in practice. An informal test of the assumption of parallel regression lines is made by transforming the ordered categorical variable into a set of dummy variables, following the cumulative probability structure. Thus, for an outcome variable with $C$ categories, $C - 1$ dummies are created. The first dummy variable equals 1 if the response is in category 1, and 0 otherwise; the second dummy variable equals 1 if the response is in category 2 or 1, and 0 otherwise and so on until the last dummy variable, which equals 1 if the response is in category $C - 1$ or lower, and 0 if the response is in category $C$. Finally, independent regressions are carried out on all dummies, and the null-hypothesis of equal regression coefficients is informally assessed by inspecting the estimated regression coefficients and their standard errors. Long (1997) gives an example of this procedure and describes a number of formal statistical tests.

There are a number of alternatives for the proportional odds model. Hedeker and Mermelstein (1998) describe a multilevel model that relaxes the proportional odds assumption, by modeling the thresholds separately. This allows predictors to have varying effects across different cut-off points. Other approaches include adding interactions with the thresholds to the model, or analyzing the ordinal data with a multinomial model, using only the categorical nominal information. Adjacent category or continuation ratio logits are other options. These are well known in the single-level regression literature, but their extension to the multilevel case and implementation in software is limited. If there are only a few predictors that fail to satisfy the proportional odds assumption, it may be possible to use a partial proportional odds model, where most predictors do meet that assumption, but a few do not.

### 7.1.3 Example of ordered categorical data

Assume that we undertake a survey to determine how characteristics of streets affect feelings of unsafety in people walking these streets. A sample of 100 streets are selected, and on each street a random sample of 10 persons are asked how often they feel unsafe while walking that street. The question on safety is asked using three answer categories: 1 = never, 2 = sometimes, 3 = often. Predictor variables are age and gender; street characteristics are an economic index (standardized $Z$-score) and a rating of the crowdedness of the street (7-point scale). The data have a multilevel structure, with people nested in streets.

Figure 7.2 shows the distribution of the outcome variable. In addition to having only three categories, it is clear that the distribution is not symmetric. Treating this outcome as continuous is not proper.

These data have some characteristics that make estimation unnecessarily difficult. The respondents' ages are recorded in years, and range from 18 to 72. This range is very different from the range of the other variables. In addition, zero is not a possible value in either age or crowdedness. To deal with these issues, the variable age is divided by 10, and all explanatory variables (including sex) are centered. Using the exploration strategy proposed earlier (Chapter 4), it turns out that all explanatory variables are

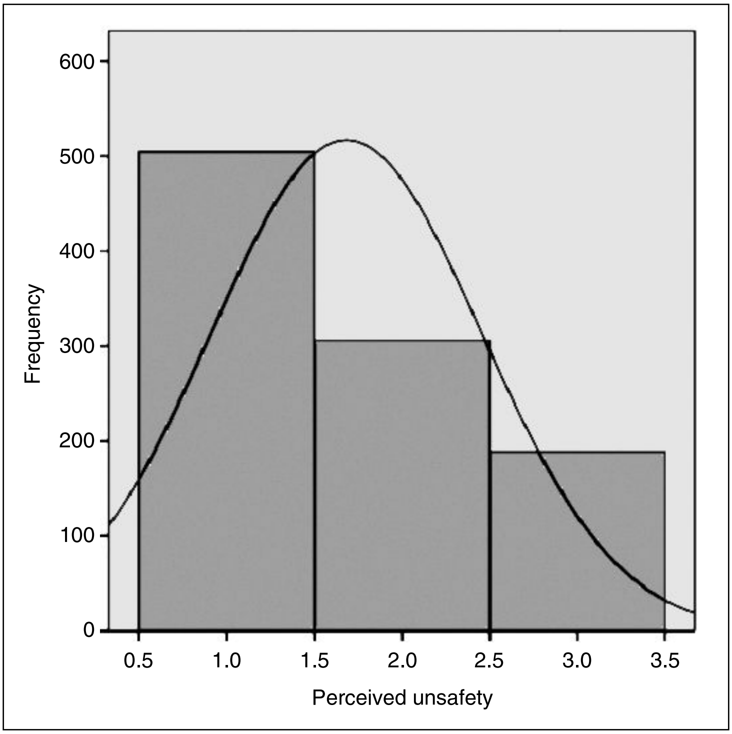


*Figure 7.2* Histogram of outcome variable for feeling unsafe.

significant, and that age has a significant slope variation across streets. However, this variation cannot be explained by economic status or crowdedness. Table 7.1 presents the results of the final model, once for the logit and once for the probit model. Estimates were made with full ML, using numerical approximation in SuperMix.

In terms of interpretation the two models are equivalent. The coefficients and the standard errors are on average 1.7 times larger in the logit model than in the probit model. The variances and their standard errors are 2.73 times larger, which is approximately 1.65 squared. The probit model is simple to interpret, since the underlying scale has a standard deviation of 1. So, an increase in age by 10 years increases the feelings of unsafety by approximately one-fourth of a standard deviation, which is a relatively small effect. On the other hand, the difference between men and women on the underlying scale is about 0.7 standard deviation, which is a large effect. On the logit scale, the interpretation is often in terms of the odds ratio. Thus, the odds ratio corresponding to the regression coefficient of 0.46 for age/10 is $e^{0.46} = 1.59$. Thus, a difference of 10 years results in an odds ratio for being in response category $c$ compared to $c - 1$ that is 1.59

*Table 7.1* Results of logit and probit model for unsafety data

| Model | Logit | Probit |
|---|---|---|
| **Fixed part** | Coeff. (s.e.) | Coeff. (s.e.) |
| Intercept | –0.02 (.09) | –0.02 (.06) |
| Threshold 2 | 2.02 (.11) | 1.18 (.06) |
| Age/10 | 0.46 (.07) | 0.27 (.04) |
| Sex | 1.23 (.15) | 0.72 (.09) |
| Economic | –0.72 (.09) | –0.42 (.05) |
| Crowded | –0.47 (.05) | –0.27 (.03) |
| **Random part** | | |
| Intercept | 0.26 (.12) | 0.10 (.04) |
| Age/10 | 0.20 (.07) | 0.07 (.02) |
| Int/age | –0.01 (.07) | –0.00 (.02) |
| Deviance | 1718.58 | 1718.08 |

times larger. Note that the change in odds ratio is independent of the specific response category, which follows from the proportional odds assumption.

To gain some insight into the effect of different estimation strategies, Table 7.2 presents the same results for the logit model only, where the estimation methods are varied. The first column contains the estimates produced using Taylor series expansion (using HLM, first order PQL). The second column contains the estimates using numerical integration with SuperMix, and the third column contains the estimates using numerical integration with Mplus, which uses a different estimation algorithm.

All estimates in Table 7.2 are close. The Taylor series linearization in HLM produces estimates that are a bit smaller than those produced by the numerical integration methods. For dichotomous data it has been shown that the Taylor series approach tends to have a small negative bias (Breslow & Lin, 1995; Raudenbush, Yang, & Yosef, 2000; Rodriguez & Goldman, 1995). The estimates in Table 7.2 suggest that the same bias occurs in modeling ordered data. Nevertheless, the estimates produced by Taylor series approximation for the unsafety data are very close to the other estimates, and the differences would not lead to a different interpretation. The estimates produced by the numerical integration in SuperMix and Mplus are essentially identical. HLM does not give standard errors for the random part, but the chi-square test on the residuals (see Chapter 3 for details) shows that both the intercept and the slope variance are significant.

*Table 7.2* Results for unsafety data with different estimation methods

| Estimation | Taylor series<br>(HLM) | Numerical<br>(SuperMix) | Numerical<br>(Mplus) |
|---|---|---|---|
| **Fixed part** | Coeff. (s.e.) | Coeff. (s.e.) | Coeff. (s.e.) |
| Intercept/Thresh | –0.01 (.09) | –0.02 (.09) | 0.02 (.09) |
| Threshold 2 | 1.96 (.10) | 2.02 (.11) | 2.04 (.12) |
| Age/10 | –0.42 (.06) | 0.46 (.07) | 0.46 (.07) |
| Sex | –1.15 (.14) | 1.23 (.15) | 1.22 (.14) |
| Economic | 0.68 (.09) | –0.72 (.09) | –0.72 (.09) |
| Crowded | 0.44 (.05) | –0.47 (.05) | –0.47 (.05) |
| **Random part** | | | |
| Intercept | 0.21 | 0.26 (.12) | 0.26 (.07) |
| Age/10 | 0.16 | 0.20 (.07) | 0.20 (.07) |
| Int/Age | –0.01 | –0.01 (.07) | –.01 (.07) |
| Deviance | | 1718.58 | 1718.59 |
| AIC | | 1736.58 | 1736.59 |
| BIC | | 1780.75 | 1780.76 |

Table 7.2 also illustrates the effect of different choices for the model parameterization. HLM uses the proportional odds model as presented in equation 7.8. This models the probability of being in category *c* or lower against being in the last category *c* = *C*. Thus, the regression coefficients have a sign that is opposite to the sign in an ordinary regression model. SuperMix and Mplus use the model as presented in equation 7.10, where the linear predictor in the generalized regression model is multiplied by –1 to restore the signs of the regression coefficients. A small difference between SuperMix and Mplus is that SuperMix transforms the thresholds as described above, and Mplus does not. So the first row in the fixed part shows the intercept for SuperMix, and threshold 1 for Mplus. If we subtract 0.02 from both thresholds in the Mplus column, the first becomes 0 and the second becomes identical to threshold 2 in the SuperMix column. All these model parameterizations are equivalent, but the opposite signs of the regression coefficients in Table 7.2 show the importance of knowing exactly what the software at hand actually does.

## 7.2 COUNT DATA

Frequently the outcome variable of interest is a count of events. In most cases count data do not have a nice normal distribution. A count can not be lower than zero, so count data always have a lower bound at zero. In addition, there may be a number of extreme values, which often results in a long tail at the right and hence skewness. When the outcome is a count of events that occur frequently, these problems can be addressed by taking the square root or in more extreme cases the logarithm. However, such nonlinear transformations change the interpretation of the underlying scale, so analyzing counts directly may be preferable. Count data can be analyzed directly using a generalized linear model. When the counted events are relatively rare they are often analyzed using a Poisson model. Examples of such events are frequency of depressive symptoms in a normal population, traffic accidents on specific road stretches, or conflicts in stable relationships. More frequent counts are often analyzed using a negative binomial model. Both models will be presented in the next section.

### 7.2.1 The Poisson model for count data

In the Poisson distribution, the probability of observing $y$ events ($y = 0, 1, 2, 3, \ldots$) is:

$$\Pr(y) = \frac{\exp(-\lambda)\lambda^{y}}{y!}, \tag{7.11}$$

where *exp* is the inverse of the natural logarithm. Just like the binomial distribution, the Poisson distribution has only one parameter, the event rate $\lambda$ (lambda). The mean and variance of the Poisson distribution are both equal to $\lambda$. As a result, with an increasing event rate, the frequency of the higher counts increases, and the variance of the counts also increases, which introduces heteroscedasticity. An important assumption in the Poisson distribution is that the events are independent and have a constant mean rate ($\lambda$). For example, counting how many days a pupil has missed school is probably not a Poisson variate, because one may miss school because of an illness, and if this lasts several days these counts are not independent. The number of typing errors on randomly chosen book pages is probably a Poisson variate.

The Poisson model for count data is a generalized linear regression model that consists of three components:

1. an outcome variable $y$ with a specific error distribution that has mean $\mu$ and variance $\sigma^2$,
2. a linear additive regression equation that produces a predictor $\eta$ of the outcome variable $y$,

3. a *link function* that links the expected values of the outcome variable $y$ to the predicted values for $\eta$: $\eta = f(\mu)$.

For counts, the outcome variable is often assumed to follow a Poisson distribution with event rate $\lambda$. The Poisson model assumes that the length of the observation period is fixed in advance (constant exposure), the events occur at a constant rate, and that the number of events at disjoint intervals are statistically independent. The multilevel Poisson model deals with certain kinds of dependence. The model can be further extended by including a varying exposure rate $m$. For instance, if book pages have different numbers of words, the distribution of typing errors would be Poisson, with exposure rate the number of words on a page. In some software the exposure variable just needs to be specified. If this is not possible, the exposure variable is added as a predictor variable to the model, including a log transformation ln($m$) to put it on the same scale as the latent outcome variable $\eta$. Such a term is called the *offset* in the linear model, and usually its coefficient is constrained equal to 1 (McCullagh & Nelder, 1989).

The multilevel Poisson regression model for a count $Y_{ij}$ for person $i$ in group $j$ can be written as:

$$Y_{ij}|\lambda_{ij} = \text{Poisson}\ (m_{ij}, \lambda_{ij}). \tag{7.12}$$

The standard link function for the Poisson distribution is the logarithm, and:

$$\eta_{ij} = \ln\ (\lambda_{ij}). \tag{7.13}$$

The level 1 and level 2 models are constructed as usual, so:

$$\eta_{ij} = \beta_{0j} + \beta_{1j}X_{ij}, \tag{7.14}$$

and:

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01}Z_j + u_{0j}\\ \beta_{1j} &= \gamma_{10} + \gamma_{11}Z_j + u_{1j},\end{aligned} \tag{7.15}$$

giving:

$$\eta_{cij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + \gamma_{11}X_{ij}Z_j + u_{0j} + u_{1j}X_{ij}. \tag{7.16}$$

Since the Poisson distribution has only one parameter, specifying an expected count implies a specific variance. Hence, the first-level equations do not have a lowest-level error term. In actual practice, we often find that the variance exceeds its expected value. In this case we have overdispersion. In more rare cases we may have underdispersion. Underdispersion often indicates a misspecification of the model, such as the omission of large interaction effects. Overdispersion can occur if there are extreme outliers, or if we omit an entire level in a multilevel model. In binomial models, very small group sizes (around three or less) also lead to overdispersion (Wright, 1997); this is likely to be also the case in Poisson models. A different problem is the problem of having many more zero counts than expected. This problem is dealt with later in this chapter.

*Example of count data*

Skrondal and Rabe-Hesketh (2004) discuss an example where 59 patients who suffer from epileptic seizures are followed on four consecutive visits to the clinic. There is a baseline count of the number of epileptic seizures in the 2 weeks before the treatment starts. After the baseline count, the patients are randomly assigned to a treatment (drug) and a control (placebo) condition. One additional variable is the patients' age. Because the link function is the logarithm, the baseline seizure count and age are log-transformed, and subsequently centered around their grand mean.

Figure 7.3 shows the frequency distribution of the seizures, which is evidently not normally distributed. The mean number of seizures is 8.3 and the variance is 152.7, which casts serious doubt on the applicability of the Poisson model. However, the histogram also shows some extreme outliers. Skrondal and Rabe-Hesketh (2004) discuss these data in more detail, pointing out how inspection of residuals and related procedures provides information about the model fit.

Table 7.3 presents the results of a multilevel Poisson regression analysis of the epilepsy data. We omit the age variable, which is not significant. The trend over time, represented by the regression coefficient for the variable 'visit', is significant but very small, and therefore also excluded. Consequently, the lowest level contains just the four counts of epileptic seizures as multivariate outcomes, without modeling a structure over time. Given the heterogeneity in the data, robust standard errors are used where available. HLM does not give standard errors for the random part, but the chi-square test on the residuals (see Chapter 3 for details) shows that the intercept variance is significant. The three different analyses result in very similar estimates. A fourth analysis with SuperMix (not presented here), which uses a different type of numerical approximation than HLM, results in virtually the same estimates as the HLM numerical approach presented in the last column. Note that HLM does not allow overdispersion with the numerical approximation. SuperMix also does not allow an overdispersion parameter in the model, but it too can estimate models for count data using a negative binomial model. This is an extension of the Poisson model that allows extra variance in the counts.

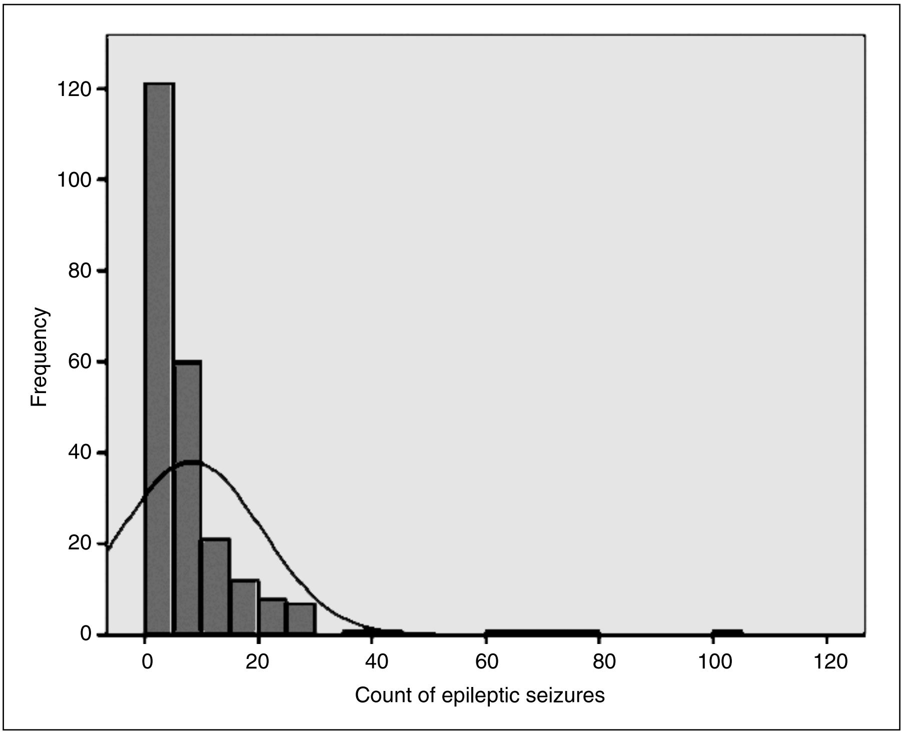


*Figure 7.3* Frequency distribution of epileptic seizures.

*Table 7.3* Results for epilepsy data with different estimation methods

| Estimation | Taylor series (HLM) | Taylor series (HLM) | Numerical (HLM) |
|---|---|---|---|
| **Fixed part** | Coeff. (s.e.)$^r$ | Coeff. (s.e.)$^r$ | Coeff. (s.e.)$^r$ |
| Intercept | 1.82 (.09) | 1.83 (.09) | 1.80 (.13) |
| Log baseline | 1.00 (.10) | 1.00 (.11) | 1.01 (.10) |
| Treatment | –0.33 (.15) | –0.30 (.15) | –0.34 (.16) |
| **Random part** | | | |
| Intercept | 0.28 | 0.26 | 0.27 |
| Overdispersion | | 1.42 | |

$^r$ Indicates robust standard error used.

With all estimation methods the baseline measurement has a strong effect, and the treatment effect is significant at the .05 level. To interpret the results in Table 7.3, we need to translate the estimates on the log scale to the observed events. The log baseline is centered, and the control group is coded 0, so the intercept refers to the expected event rate for the control group. Using the estimates in the last column, we take $Y = e^{1.8} = 6.05$ as the event rate in the control group. In the experimental group we take $Y = e^{(1.8 - 0.34)} = 4.31$ as the event rate. On average, the drug lowers the event rate by 28.8% of the event rate in the untreated control group.

### 7.2.2 The negative binomial model for count data

In the Poisson model, the variance of the outcome is equal to the mean. When the observed variance is much larger than expected under the Poisson model, we have overdispersion. One way to model overdispersion is to add an explicit error term to the model. Thus, for the Poisson model we have the link function (see 7.13) $\eta_{ij} = \ln(\lambda_{ij})$, and the inverse is $\lambda_{ij} = \exp(\eta_{ij})$, where $\eta_{ij}$ is the outcome predicted by the linear regression model. The negative binomial adds an explicit error term $\varepsilon$ to the model, as follows:

$$\lambda_{ij} = \exp(\eta_{ij} + \varepsilon_{ij}) = \exp(\eta_{ij})\exp(\varepsilon_{ij}) \tag{7.17}$$

The error term in the model increases the variance compared to the variance implied by the Poisson model. This is similar to adding a dispersion parameter in a Poisson model, and a detailed description of the single-level negative binomial model is given by Long (1997). When the epilepsy data are analyzed with the negative binomial model, the estimates are very close to the results in the last column of Table 7.3. The dispersion parameter is 0.14 (s.e. = .04, $p$ = .001). The subject-level variance is a bit lower at 0.24, which is reasonable given that in the negative binomial model there is more variance at the event level than in the Poisson model presented in Table 7.3. Given that the negative binomial model is a Poisson model with an added variance term, the test on the deviance can be used to assess whether the negative binomial model fits better. The negative binomial model cannot directly be compared to the Poisson model with overdispersion parameter, because these models are not nested. However, the AIC and BIC can be used to compare these models. Both the AIC and the BIC are smaller for the Poisson model with overdispersion than for the negative binomial model.

### 7.2.3 Too many zeros: The zero inflated model

When the data show an excess of zeros compared to the expected number under the Poisson distribution, it is sometimes assumed that there are two processes that produce the data. Some of the zeros are part of the event count, and are assumed to follow a Poisson model (or a negative binomial). Other zeros are part of the event taking place or not, a binary process modeled by a binomial model. These zeros are not part of the count, they are structural zeros, indicating that the event *never* takes place. Thus, the assumption is that our data actually include two populations, one that always produces zeros and a second that produces counts following a Poisson model. For example, assume that we study risky behavior, such as using drugs or having unsafe sex. One population never shows this behavior: it is simply not part of the behavior repertoire. These individuals will always report a zero. The other population consists of individuals who do have this behavior in their repertoire. These individuals can report on their behavior, and these reports can also contain zeros. An individual may sometimes use drugs, but just did not do this in the time period surveyed. Models for such mixtures are referred to as zero inflated Poisson or ZIP models. For the count part of the model we use a standard regression model, for instance assuming a Poisson or a negative binomial distribution, and for the probability of being in the population that can produce only zeros we use a standard logistic regression model. Both models are estimated simultaneously. Table 7.4 presents the results for a multilevel Poisson and a multilevel zero inflated Poisson (ZIP) model for the epilepsy data (using Mplus).

*Table 7.4* Results for epilepsy data: Poisson and zero inflated Poisson

| Estimation | Poisson | ZIP |
|---|---|---|
| **Fixed part** | Coeff. (s.e.)[r] | Coeff. (s.e.)[r] |
| Intercept | 1.80 (.09) | 1.87 (.09) |
| Log baseline | 1.01 (.11) | 0.99 (.11) |
| Treatment | –0.34 (.15) | –0.35 (.15) |
| Inflation intercept | | –3.08 (.49) |
| **Random part** | | |
| Intercept | 0.28 (.07) | 0.25 (.06) |
| Deviance | 1343.20 | 1320.29 |
| AIC | 1351.20 | 1330.29 |
| BIC | 1365.05 | 1347.61 |

[r] Indicates robust standard error used.

Both the AIC and the BIC indicate that the ZIP model is better, although the parameter estimates for the Poisson model change very little. There is an extra parameter: the intercept of the inflation part. In the ZIP model reported in Table 7.4, there are no explanatory variables that predict the probability of being in the always zero class. As a result, the intercept indicates the average probability of being in that class. A large value of the intercept indicates a large fraction of ‘always zero’. The model for the inflation part is a logistic model, so the intercept value of –3.08 is on the underlying logit scale. Translating it into a proportion using the inverse of the logit transformation (introduced in Chapter 6), we find:

$$\hat{p} = \frac{e^{3.08}}{1 + e^{3.08}} = .044, \tag{7.18}$$

which shows that the fraction of ‘always zero’ in the epilepsy data is very small. In the epilepsy data set, 9.7% of the subjects report zero seizures. Using 7.18, we can now estimate that 4.4% have no seizures, meaning that their epilepsy is totally suppressed, and 5.3% of the subjects merely happen to have no seizures in the period surveyed.

The ZIP model reported in Table 7.4 does not include predictors for the inflation part. It is possible to expand the inflation model, which is a logistic model similar to the models discussed in Chapter 6, by including predictors. In this particular data set, the available predictors do not predict the zero inflation, and the AIC and BIC indicate that the ZIP model without predictors for the inflation part is preferable.

Empirical data often exhibit overdispersion or an excess of zeros, and choosing the best distributional model is an important part of the modeling process (see Gray, 2005, for an example).

Just like the Poisson model, the negative binomial model can also be extended to include an inflated numbers of zeros. Lee, Wang, Scott, Yau, and McLachlan (2006) provide a discussion of the multilevel Poisson model for data with an excess of zeros, and Moghimbeigi, Esraghian, Mohammad, and McArdle (2008) discuss negative binomial models for such data. In the epilepsy example data, adding a zero inflation part to the negative binomial model turns out to be superfluous, the latent class of extra zeros is estimated as very small, and the AIC and BIC indicate that the negative binomial model without zero inflation is preferable.

## 7.3 THE EVER CHANGING LATENT SCALE, AGAIN

Just as in logistic and probit regression, the scale of the latent outcome variable implicitly changes when the model is changed. The lowest-level residuals in each separate model are scaled to the variance of the standard distribution. This variance is

$\pi^2/3 \approx 3.29$ in the logistic distribution and 1 in the normal distribution. In the Poisson distribution, the variance is equal to the (predicted) mean. In the negative binomial distribution, an extra error term is added to the Poisson variance. With some changes in calculating the lowest-level residual variance, all procedures discussed in section 6.5 also apply to the ordered and count data discussed in this chapter.

# 8

# Multilevel Survival Analysis

Survival analysis or event history analysis is a set of methods for modeling the length of time until the occurrence of some event. The term 'survival analysis' is used most often in biomedical applications where the event of interest is not repeatable, such as studies where the length of time until relapse or loss of life is modeled as a function of some independent variables, often including a treatment indicator. Failure analysis and hazard modeling are related terms. In social research, more neutral terms like event history analysis or duration analysis are commonly used. This chapter uses the term survival analysis to refer to all these methods. An important feature of survival data is that for some cases the final event is not yet observed, and such observations are said to be censored. This chapter provides a brief introduction to survival analysis, and shows how standard survival analysis can be related to multilevel analysis. Next, it discusses the multilevel analysis of survival data.

## 8.1 SURVIVAL ANALYSIS

Survival analysis concerns the analysis of the occurrence of events over time and of the length of time it takes for their occurrence. One way to collect the necessary data is a longitudinal survey, where individuals are surveyed repeatedly and questioned about the occurrence of certain events, for example marriage, births, getting a job, or dropping out of school. More often, the data are collected in a retrospective survey, where respondents are asked if and at what time certain events took place. Sometimes events may occur more than once, for instance multiple marriages may occur, or a woman may have several children. This is often denoted as hazard modeling, where the hazard of the event is modeled as a function of time since the last occurrence.

A central issue in survival or event history data is that observations may arise where the occurrence of an event has not yet been observed, because the subject left the study or because the study ended at a point in time before the event could be observed. For these subjects, the survival time is unknown; these observations are said to be censored. More precisely, these observations are right-censored: the event has not yet occurred at the end of the observation period. Simply excluding these observations is not an acceptable solution. Imagine that in a longitudinal study of marriage duration we exclude all respondents that are still married at the end of the data collection period. Estimates of average marriage duration would obviously be biased downward, and we would have discarded a large fraction of our sample. Thus, the goal of survival analysis is to obtain unbiased estimates of expected duration, including relevant predictor variables, and incorporating censored observations.

Survival analysis requires two separate elements: a variable with the time information and a variable that indicates the censoring status. The time variable provides the duration of time until the event, or until the last observation, and the censoring status indicates whether the last observation is censored or not. Two general approaches can be distinguished: one that treats time as discrete and grouped, analogous to longitudinal data with fixed occasions, and one that treats time as continuous, analogous to longitudinal data with varying occasions.

The grouped time approach categorizes the time variable into a relatively small number of time intervals. For each time interval, it determines the cumulative proportion of subjects that survive to the beginning of that interval (the survival rate), and how many subjects who enter the interval experience the event before the midpoint of that interval (the hazard). The survival rate and the hazard can be summarized in life tables, and various statistics are available to summarize the survival and hazard for the entire group or for subgroups. The survival function, the cumulative proportion of cases surviving to the beginning of the $t + 1$ interval, is:

$$P_{t+1} = p_t P_t, \tag{8.1}$$

or simply the proportion who survived until the start of the interval $P_t$ multiplied by the proportion $p_t$ of subjects at risk that survived the interval $t$ (see Tabachnick & Fidell, 2007). The probability of experiencing the event in interval $t$ is equal to the number of events in that interval divided by the number at risk. The discrete time estimator of the hazard rate is then:

$$h(t) = \frac{events_t}{at\ risk_t}. \tag{8.2}$$

Unless the duration data are actually collected at regular fixed occasions, treating time as a discrete fixed variable appears to be an inferior approach. Several approaches to continuous duration data have been developed, of which the Kaplan-Meier method and the Cox method are most widely used. The Kaplan-Meier method uses the grouped time method, but extends this by putting each observed duration value into its own interval. Next, the same methods are applied as in the ordinary grouped time approach. The appropriate techniques are treated in detail in Singer and Willett (2003).

The survival model is basically a model for the hazard function. The hazard function reflects the probability of experiencing the event in time interval $t$ given that the event was not previously experienced. The hazard is modeled by covariates $X_{ti}$ which can be time-invariant or time-varying. Since it is a probability, the hazard function is bounded between zero and one. Therefore, a linear model for the hazard is not suitable, and a generalized linear model is used, with an appropriate link function to transform the hazard. For a given link function $g$ we have:

$$g(h(t \mid X_{ti})) = \alpha_t + \beta_1 X_{ti}, \tag{8.3}$$

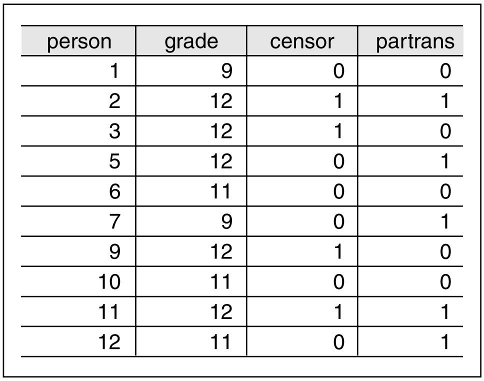
where $\alpha_t$ are time-specific intercepts that represent the baseline hazard. Thus, each interval has its own intercept $\alpha_t$. If the link function $g$ is the logit function, the corresponding model is the proportional odds model:

$$\ln\left(\frac{h(t \mid X_{ti})}{1 - h(t \mid X_{ti})}\right) = \alpha_t + \beta_1 X_{ti}. \tag{8.4}$$

Other link functions used in survival modeling are the log function and the complementary log-log function given by $g = -\ln(-\ln(1 - \text{p}))$. In the proportional odds model given by 8.4, a unit change in covariate $X_{ti}$ produces a $\beta_1$ change in the logit of the hazard, and this effect is independent of the time interval. If there are a large number of distinct time points, there are as many time-specific intercepts $\alpha_t$, which leads to a large model with many parameters. To make the model more parsimonious, in such cases the time-specific intercepts are approximated by a function of the time variable $t$, for instance a high-degree polynomial. Since $T$ distinct time points can be fitted perfectly by a polynomial of degree $T - 1$, the model with a polynomial of degree $R < T - 1$ is nested in the model with $T$ time-specific intercepts, and the chi-square difference test on the deviance can be used to test whether a polynomial of degree $R$ is sufficient. Thus, the $T$ time-specific intercepts are replaced by a regression equation of the form:

$$\alpha_t = \gamma_0 + \gamma_1 t + \ldots \gamma_{T-1} t^{T-1}, \tag{8.5}$$

As an example we use a data set collected by Capaldi, Crosby, and Stoolmiller (1996) and discussed in depth by Singer and Willet (2003). This example shows how a survival model can be viewed as a special multilevel model, after which the extension to more levels is straightforward. The data were a sample of 180 boys who were tracked from the 7th grade (approximate age 12 years) until the 12th grade (approximate age 17). The event of interest was time to first sexual intercourse; in the 12th grade 30% of the boys were still virgins, meaning that these observations are censored. We used one time-invariant predictor, a dummy variable indicating whether before the 7th grade the boys had experienced a parental transition (coded 1) or had lived with both parents during these years (coded 0). The interest was in the question of whether the time to first sexual intercourse was the same for both groups.

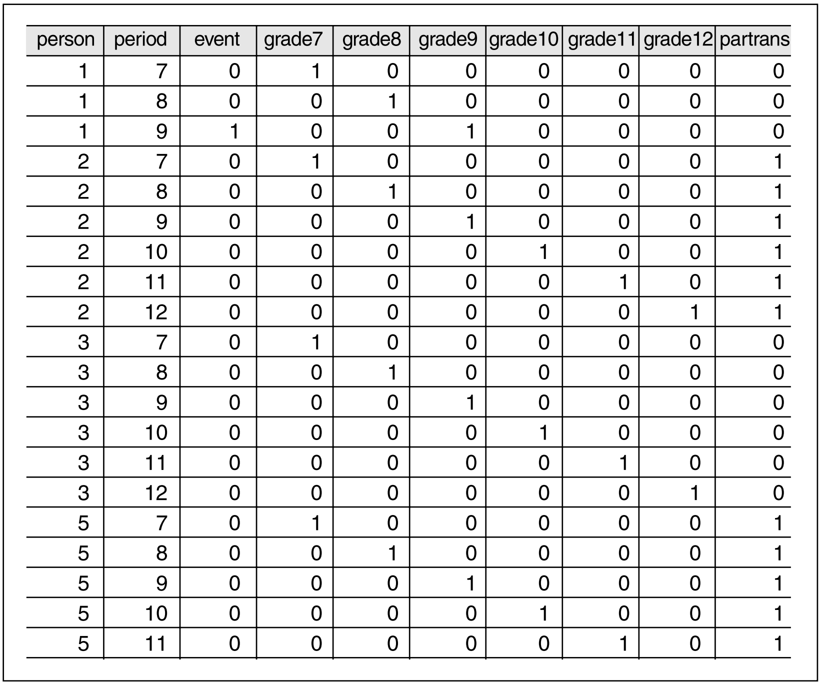


| person | grade | censor | partrans |
|---|---|---|---|
| 1 | 9 | 0 | 0 |
| 2 | 12 | 1 | 1 |
| 3 | 12 | 1 | 0 |
| 5 | 12 | 0 | 1 |
| 6 | 11 | 0 | 0 |
| 7 | 9 | 0 | 1 |
| 9 | 12 | 1 | 0 |
| 10 | 11 | 0 | 0 |
| 11 | 12 | 1 | 1 |
| 12 | 11 | 0 | 1 |

*Figure 8.1* Normal data file for survival data.

Figure 8.1 shows the data file in a standard 'normal' format, with one line of data for each subject. The variable *grade* records the grade in which the event happens, after which observation stops. If the event has not happened by the 12th grade, the grade recorded is 12, and the observation is coded as censored. The variable *partrans* indicates whether a parental transition had occurred before grade 7 or not. To analyze these data in standard logistic regression software, the data must be restructured to the 'person–period' format shown in Figure 8.2. In Figure 8.2, each row in the data corresponds to a specific person and grade combination, and there are as many rows as there are times (grades) to the event. There are several data lines for each subject, similar to the 'long' data format used in longitudinal modeling (Chapter 5). To model the baseline hazard in each grade, six dummies have been added to indicate each of the six grades.

Both equation 8.1 and the structure of the data in Figure 8.2 suggest a multilevel approach to these longitudinal data. However, multilevel modeling of within-person variation that consists of a long series of zeros followed (sometimes) by a single *one* leads to estimation problems. The data as given in Figure 8.1 can be modeled using special survival analysis software. It is also possible to use standard logistic regression software, where the dependencies in the data are dealt with by adding the period dummies to the model. If there are many periods, there will be many dummies. In that case, the baseline hazards (the period-specific intercepts) are approximated by a smooth function of time, such as a polynomial function of time.



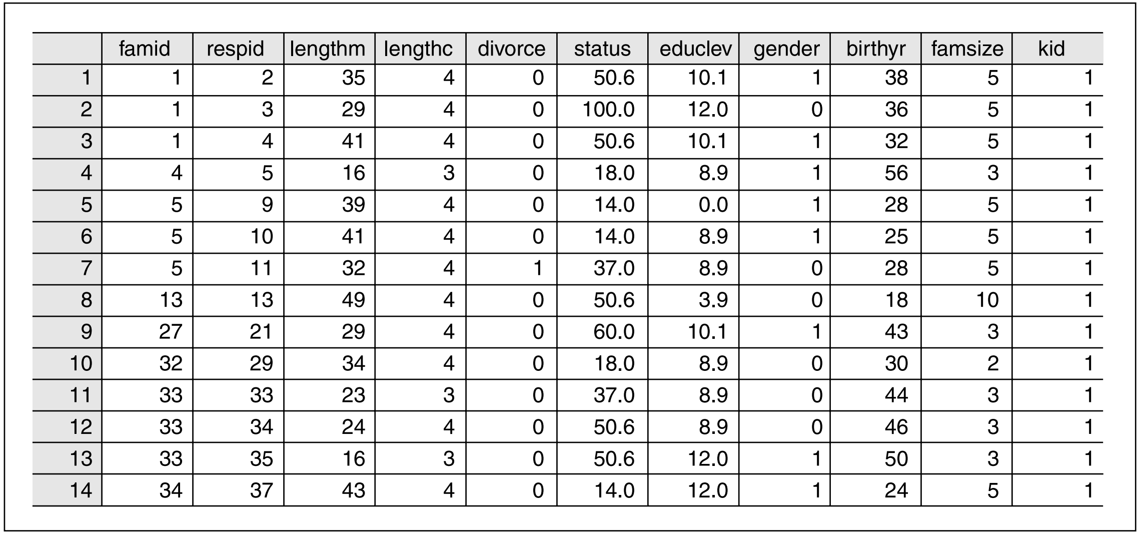
| person | period | event | grade7 | grade8 | grade9 | grade10 | grade11 | grade12 | partrans |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 7 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 8 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 1 | 9 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 2 | 7 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2 | 8 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| 2 | 9 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 2 | 10 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 2 | 11 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 2 | 12 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 3 | 7 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 8 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 3 | 9 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 3 | 10 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 3 | 11 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 3 | 12 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 5 | 7 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 5 | 8 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| 5 | 9 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 5 | 10 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 5 | 11 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |

*Figure 8.2* Person–period format of survival data.

## 8.2 MULTILEVEL SURVIVAL ANALYSIS

The discrete or grouped survival model extends readily to multilevel models (see Barber, Murphy, Axinn, & Maples, 2000; Reardon, Brennan, & Buka, 2002; and Grilli, 2005), and this chapter restricts itself to an exposition of this model. As an example, we use a data set about divorce risk, which was analyzed earlier by Dronkers and Hox (2006). The data are described in more detail in Appendix A. The data set consists of longitudinal data where respondents were asked repeatedly about live events, such as marriages, divorces, and child births. In addition to their own history, respondents were also asked to provide information on up to three siblings. The data set analyzed here includes only data on the first marriage.

Figure 8.3 shows part of the data. There are respondents nested within families. There is a duration variable *lengthm* (length of marriage in years), which indicates the time when the last observation is made, and a variable indicating if the subject has divorced. If the last observed value for divorced equals zero, that observation is censored. Other variables indicate the respondent's occupational status, educational level,



| | famid | respid | lengthm | lengthc | divorce | status | educlev | gender | birthyr | famsize | kid |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 2 | 35 | 4 | 0 | 50.6 | 10.1 | 1 | 38 | 5 | 1 |
| 2 | 1 | 3 | 29 | 4 | 0 | 100.0 | 12.0 | 0 | 36 | 5 | 1 |
| 3 | 1 | 4 | 41 | 4 | 0 | 50.6 | 10.1 | 1 | 32 | 5 | 1 |
| 4 | 4 | 5 | 16 | 3 | 0 | 18.0 | 8.9 | 1 | 56 | 3 | 1 |
| 5 | 5 | 9 | 39 | 4 | 0 | 14.0 | 0.0 | 1 | 28 | 5 | 1 |
| 6 | 5 | 10 | 41 | 4 | 0 | 14.0 | 8.9 | 1 | 25 | 5 | 1 |
| 7 | 5 | 11 | 32 | 4 | 1 | 37.0 | 8.9 | 0 | 28 | 5 | 1 |
| 8 | 13 | 13 | 49 | 4 | 0 | 50.6 | 3.9 | 0 | 18 | 10 | 1 |
| 9 | 27 | 21 | 29 | 4 | 0 | 60.0 | 10.1 | 1 | 43 | 3 | 1 |
| 10 | 32 | 29 | 34 | 4 | 0 | 18.0 | 8.9 | 0 | 30 | 2 | 1 |
| 11 | 33 | 33 | 23 | 3 | 0 | 37.0 | 8.9 | 0 | 44 | 3 | 1 |
| 12 | 33 | 34 | 24 | 4 | 0 | 50.6 | 8.9 | 0 | 46 | 3 | 1 |
| 13 | 33 | 35 | 16 | 3 | 0 | 50.6 | 12.0 | 1 | 50 | 3 | 1 |
| 14 | 34 | 37 | 43 | 4 | 0 | 14.0 | 12.0 | 1 | 24 | 5 | 1 |

*Figure 8.3* Divorce example data (part).

gender (female = 1), parental family size, whether the respondent has kids, educational level of mother and father, and the proportion of siblings that are divorced. The research problem involved the question whether divorce risks are similar for children from the same family (siblings) and whether parental characteristics can explain the similarity of divorce risk.

To analyze these data using the discrete time survival model, the data must be restructured to a format resembling the stacked or 'long' format used to analyze longitudinal data with multilevel models. The result is a person–period data set with one row for each person–period combination. Figure 8.3 shows part of the person–period data for some of the respondents shown in Figure 8.4.

In Figure 8.4 we see part of the restructured data set; the last four observations for respondent 2 (who was married for 35 years at the interview time, and had not divorced) and the first six observations for respondent 3. So the last value of the variable 'divorce' for each respondent indicates either the event (scored 1) or censoring (scored 0).

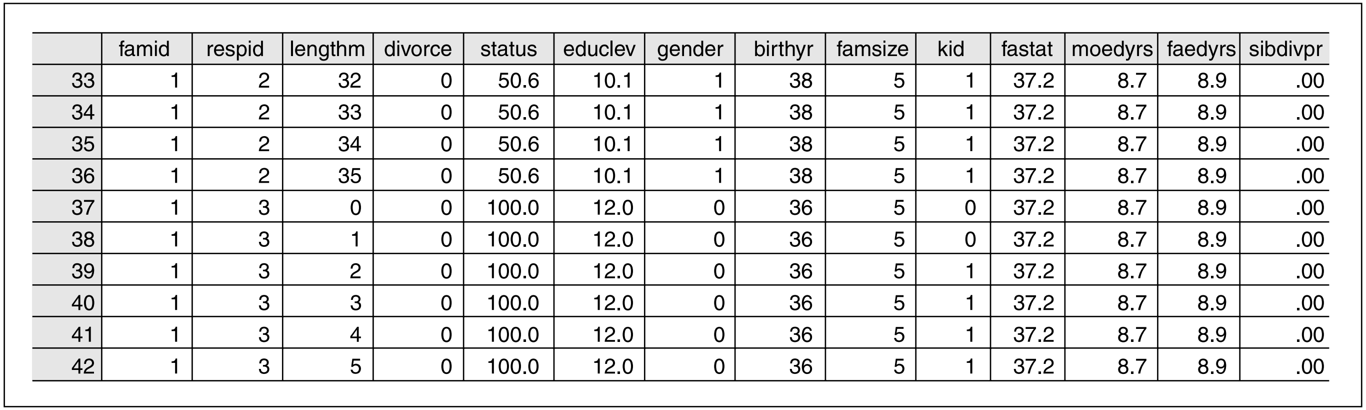
To estimate the $t$ time-specific intercepts we need to add $t$ dummy variables as period indicators to the model. The time indicator 'lengthm' for marriage length ranges in value from 0 to 67. If there are many periods the time-specific intercepts $\alpha_t$ are usually replaced by a smooth function, for instance a polynomial function of the time variable. Time-varying effects can be entered by allowing interactions of covariates with the time variable. It is clear that if the data structure in Figure 8.4 is analyzed using a multilevel model, adding time-varying covariates is straightforward. Furthermore, if respondents are grouped, as we have here, adding a third level for the groups is also straightforward. Note that in the logistic regression model used, there is no lowest-level error term (see Chapter 6 for details), and for simplicity the lowest level for the measures until the event is not shown in subsequent equations.

The hazard function $h(t)$ is the probability of the event occurring in the time interval $t$, conditional on no earlier occurrence. In our case, the time variable is the marriage length at time $t$. The hazard is generally modeled with a logistic regression, in our case a multilevel regression of the following form:

$$\text{logit}(h_{ij}(t)) = \alpha(t) + \beta_{2j}x_{ij} + \beta_3 z_j, \tag{8.6}$$

where $\alpha(t)$ is the baseline hazard at marriage year $t$, $x_{ij}$ is a sibling-level predictor, and $z_j$ is a family-level predictor. The regression coefficient of $x_{ij}$ varies at the family level. There is no intercept in the model, since the $\alpha(t)$ are a full set of dummy variables that model the baseline hazard. When there are a large number of periods, the model can be made more parsimonious by modeling the hazard as a smooth function of $t$. If a linear function of time suffices, we have:

$$\text{logit}(h_{ij}(t)) = \beta_{0ij} + \beta_1 t_{ijt} + \beta_{2j}x_{ij} + \beta_3 z_j, \tag{8.7}$$



| | famid | respid | lengthm | divorce | status | educlev | gender | birthyr | famsize | kid | fastat | moedyrs | faedyrs | sibdivpr |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 33 | 1 | 2 | 32 | 0 | 50.6 | 10.1 | 1 | 38 | 5 | 1 | 37.2 | 8.7 | 8.9 | .00 |
| 34 | 1 | 2 | 33 | 0 | 50.6 | 10.1 | 1 | 38 | 5 | 1 | 37.2 | 8.7 | 8.9 | .00 |
| 35 | 1 | 2 | 34 | 0 | 50.6 | 10.1 | 1 | 38 | 5 | 1 | 37.2 | 8.7 | 8.9 | .00 |
| 36 | 1 | 2 | 35 | 0 | 50.6 | 10.1 | 1 | 38 | 5 | 1 | 37.2 | 8.7 | 8.9 | .00 |
| 37 | 1 | 3 | 0 | 0 | 100.0 | 12.0 | 0 | 36 | 5 | 0 | 37.2 | 8.7 | 8.9 | .00 |
| 38 | 1 | 3 | 1 | 0 | 100.0 | 12.0 | 0 | 36 | 5 | 0 | 37.2 | 8.7 | 8.9 | .00 |
| 39 | 1 | 3 | 2 | 0 | 100.0 | 12.0 | 0 | 36 | 5 | 1 | 37.2 | 8.7 | 8.9 | .00 |
| 40 | 1 | 3 | 3 | 0 | 100.0 | 12.0 | 0 | 36 | 5 | 1 | 37.2 | 8.7 | 8.9 | .00 |
| 41 | 1 | 3 | 4 | 0 | 100.0 | 12.0 | 0 | 36 | 5 | 1 | 37.2 | 8.7 | 8.9 | .00 |
| 42 | 1 | 3 | 5 | 0 | 100.0 | 12.0 | 0 | 36 | 5 | 1 | 37.2 | 8.7 | 8.9 | .00 |

*Figure 8.4* Divorce example data (part).

which becomes:

$$\text{logit}(h_{ij}(t)) = \gamma_0 + \gamma_1 t_{ijt} + \gamma_2 x_{ij} + \gamma_3 z_j + u_{0j} + u_{2j}, \tag{8.8}$$

where $u_{0j}$ is the family-level variation in the overall risk, and $u_{2j}$ implies family-level variation in the slope of the sibling-level predictor variable. The regression coefficient of the period indicator $t$ may or may not vary across families, but family-level-varying baseline hazards are difficult to interpret, so the preference is to model the period indicator as fixed.

In our divorce data, the marriage length ranges from 0 to 67. Including 68 dummy variables in the model is not an attractive approach, and therefore we include a polynomial for the marriage length. The polynomial approximation must be accurate, and in addition to the linear function presented in equation 8.8 we examine polynomials up to the 5th degree polynomial. Whether these are all needed can be checked by examining their significance using the Wald test, or by a deviance difference test using full maximum likelihood and numerical approximation for the estimation.

To prevent estimation problems caused by large differences in the scale of the explanatory variables, the marriage length is transformed into a $Z$-score, and the higher degree polynomials are derived from this $Z$-score. For the divorce data, a cubic polynomial turns out to be sufficient. To evaluate how well the polynomial describes the baseline hazard, we can plot the predicted and observed proportions of divorce at each marriage length. This plot (not shown here) reveals that there is an outlier; at a marriage length of 53 years there is an unexpectedly high proportion of divorces. The number of cases with such long marriages is small (11 subjects, 0.6% of the sample), and this high divorce rate may well be chance. Still, it is best to add to the baseline hazard model a dummy that indicates period 53.

Table 8.1 shows the results for both models, with and without the dummy for marriage length equal to 53. When the period = 53 dummy is added to the model, the third order polynomial is no longer significant, and it is removed from the model. The differences between both models are not large. We have used full maximum likelihood and numerical integration (HLM Laplace method), so the deviance statistic can be used to compare the models. The models are not nested, but we can use the AIC or the BIC to compare the models. The model that treats period 53 as an outlier performs slightly better, so that will be the basis for further modeling.

The regression coefficients for time period in Table 8.1 show that the risk of divorce decreases when the marriage length is longer. The negative regression coefficient for $t^2$ indicates a faster decreasing trend for longer marriage durations. The variance of 0.58 can be compared to the lowest-level variance, which is the variance of the standard logistic distribution: $\pi^2/3 \approx 3.29$. The intraclass correlation for the

*Table 8.1* Survival model for marriage length (M1)

| Predictor | Time third order<br>Coefficient (s.e.) | Time second order + p53<br>Coefficient (s.e.) |
|---|---|---|
| Intercept | –5.71 (.16) | –5.70 (.16) |
| $t$ | –0.44 (.16) | –0.24 (.10) |
| $t^2$ | –0.46 (.13) | –0.40 (.11) |
| $t^3$ | 0.15 (.07) | – |
| Period 53 | – | 5.12 (1.34) |
| Variance $\sigma^2_{u0}$ | 0.58 ($\chi^2_{(952)} = 1394.8, p < .001$) | 0.58 ($\chi^2_{(952)} = 1354.5, p < .001$) |
| Deviance | 82347.6 | 82343.9 |
| AIC / BIC | 82357.6 / 82381.9 | 82353.9 / 82378.2 |

family-level variance in divorce risk is therefore 0.58/(3.29 + 0.58) = .15. This is not very large, but it clearly shows a family effect on divorce.

The results in Table 8.1 are best regarded as the equivalent of the intercept-only null model. Only after the person–year trend is represented well can other predictor variables be added to the model. These variables can be both time-varying and time-invariant. In our case, there are no time-varying predictors. As it turns out, none of the family-level predictors has a significant effect.

Table 8.2 presents the final model for divorce, which contains only individual-level

*Table 8.2* Survival model for marriage length (M2)

| Predictor | Time second order + p53<br>Coefficient (s.e.) | + first level explanatory vars<br>Coefficient (s.e.) |
|---|---|---|
| Intercept | –5.70 (.16) | –5.72 (.16) |
| $t$ | –0.24 (.10) | –0.21 (.10) |
| $t^2$ | –0.40 (.11) | –0.40 (.11) |
| Period 53 | 5.12 (1.34) | 5.10 (1.34) |
| EducLev | | 0.11 (.03) |
| Gender | | 0.38 (.15) |
| Variance $\sigma^2_{u0}$ | 0.58 ($\chi^2_{(952)} = 1354.5, p < .001$) | 0.52 ($\chi^2_{(952)} = 1520.7, p < .001$) |
| Deviance | 82343.9 | 82320.4 |
| AIC/BIC | 82353.9 / 82378.2 | 82334.4 / 82368.4 |

predictors. Divorce risk still decreases with marriage length, and women and respondents with a high educational level have a higher risk of divorce.

## 8.3 MULTILEVEL ORDINAL SURVIVAL ANALYSIS

Hedeker and Gibbons (2006) describe multilevel survival analysis models where the data are grouped in a small number of ordinal coded time intervals. The model assumes that there are a small number of measurement periods, coded $t = 1, 2, \ldots, T$. For each level 1 unit, observations are made until either the event occurs or censoring takes place (meaning the event had not occurred by the last observed period). As a result, we do not know the exact time point when an event has taken place, we just know in which time interval (observed period) it has taken place. The model is similar to the one in 8.6:

$$\text{logit}(h_{ij}(t)) = \alpha_t + \beta_{2j}x_{ij} + \beta_3 z_j, \tag{8.9}$$

but in 8.9 there are only a small number of period-specific intercepts, which are modeled with an ordinal threshold model, similar to the ordinal models discussed in Chapter 7. Hedeker and Gibbons (2006) express a preference for the complementary log-log function instead of the logit, which gives:

$$\ln[-\ln(1 - h_{ij}(t))] = \alpha_t + \beta_{2j}x_{ij} + \beta_3 z_j. \tag{8.10}$$

When the complementary log-log function is used, the regression coefficients in the ordinal grouped-time model are invariant to the length of the interval, and equivalent to the coefficients in the underlying continuous-time proportional hazards model. This does not hold when the logit function is used.

The advantage of the ordinal survival model is that we do not need a long series of person–period records in the data file: it suffices to have a single record for each individual, including the last observed interval and an indicator variable that specifies if the event occurred in this interval or not. In the latter case, the last observation is censored. A disadvantage of the ordinal survival model is that it can not accommodate time-varying predictors. If time-varying predictors are present, the person–period approach discussed earlier must be used.

Figure 8.3 in the previous section shows the data file for the multilevel ordered survival model, for the divorce example data analyzed earlier. Each individual respondent is present in the data file only once. In addition to the continuous variable *lengthm*, which codes the marriage length in years, there is a categorical variable *lengthc* that codes the marriage length in four categories (originally five quintiles, but the fourth and fifth quintile are combined because the fifth quintile contained only five divorces). There are up to three respondents in each family.

Table 8.3 shows the parameter estimates for the null model and for the model with the same explanatory variables used earlier, applying a logistic model estimated with SuperMix. The null model estimates the family-level variance at 0.88. The residual variance on the standard logistic distribution is fixed at $\pi^2/3 \approx 3.29$, so the intraclass correlation is .21. This estimate is somewhat higher than the estimate in the continuous grouped-time model, where the intraclass correlation was estimated as .15. It confirms the evidence for a family effect on risk of divorce.

*Table 8.3* Survival model for marriage length (M3)

| Predictor | Null model<br>Coefficient (s.e.) | + first level explanatory vars<br>Coefficient (s.e.) |
|---|---|---|
| Intercept | –3.12 (.17) | –4.33 (.37) |
| EducLev | | 0.09 (.03) |
| Gender | | 0.39 (.15) |
| Variance $\sigma^2_{u0}$ | 0.88 (.34) | 0.82 (.33) |
| Deviance | 1998.8 | 1980.4 |
| AIC/BIC | 2008.8 / 2036.4 | 1994.4 / 2033.1 |

In the null model, the thresholds are estimated as –3.12, –2.42, –2.06, and –1.84. These are the baseline hazards for the four time periods. These can be transformed to probabilities using the inverse of the logit transformation, which is:

$$P(y) = e^y/(1 + e^y). \tag{8.11}$$

Using 8.11 we calculate the baseline hazards as 0.04, 0.08, 0.11, and 0.14. The cumulative probability of experiencing a divorce increases with increasing marriage length, but it remains relatively low. As in the earlier analysis, educational level and being female increase the risk of divorce.

# 9

# Cross-Classified Multilevel Models

Not all multilevel data are purely hierarchical. For example, pupils can be nested within the schools they attend, and within the neighborhoods in which they live. However, the nesting structure may be less clear when we consider the schools and neighborhoods. It is likely that there is a tendency for pupils to attend schools in their own neighborhood, but there will be exceptions. Some pupils will attend schools in other neighborhoods than the one they live in, and, in particular, schools that are located close to the border of two neighborhoods may be expected to draw pupils from several neighborhoods. As a result, it is not possible to set up an unambiguous hierarchy of pupils within schools within neighborhoods. We could, of course, arbitrarily set up such a model structure, but to do so we would have to include several schools more than once, because they appear in several different neighborhoods.

Whenever we encounter this kind of problem, chances are that we are dealing with a cross-classified data structure. In the schools and neighborhoods example, pupils are nested within schools, and also within neighborhoods, but schools and neighborhoods are *crossed* with each other. If we study educational achievement, we may assume that achievement can be influenced by both schools and neighborhoods. Therefore, the model has to incorporate both schools and neighborhoods as sources of variation in achievement, but in such a manner that pupils are nested in the cross-classification of both schools and neighborhoods.

Cross-classified data structures can occur at any level of a hierarchical data set. If we have pupils nested within the cross-classification of schools and neighborhoods, the schools and neighborhoods are the second level, and the pupils are the lowest (first) level. However, it is also possible to have a cross-classification at the lowest level. Consider the example of students who have to carry out a set of complicated analysis tasks in a computer class. There are several parallel classes, taught by different teachers. To keep the grading policy equivalent for all students, all computer exercises are graded by all available teachers. As a result, at the student level we would have grades for several different exercises, given by several different teachers. One way to view this is to distinguish the class level as the highest level, with students as the level below this, teachers as the next lower level, and the exercises as the lowest level, below the teachers. This constitutes a nicely hierarchical four-level data structure. On the other hand, we could also distinguish the class level as the highest level, with students below this level, the exercises as the next lower level, and the teachers as the lowest level, below the exercises. This also constitutes a four-level hierarchical data structure. It appears that we can model such data using two contradictory data structures. Again, whenever we have this kind of problem, it indicates that we are probably dealing with a cross-classified data structure. In the grading example, we have pupils nested within classes, with a cross-classification of teachers (graders) and exercises nested within the pupils.

Since we may expect differences between classes and pupils, this cross-classification of exercises and teachers would be defined at the lowest level, nested within pupils within classes. The reliability of the combined grade of the student in such situations can be modeled using generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). To assess the generalizability of the students' combined grade across exercises and teachers using generalizability theory, we must partition the total variation of the assigned grades as the sum of contributions from classes, students, exercises, and teachers. Cross-classified multilevel analysis is a good way to obtain the required estimates for the various variance components in such a partition (see Hox & Maas, 2006).

Cross-classified multilevel models are applicable to a variety of situations. The examples given so far are from the field of education. Other applications are models for nonresponse in longitudinal research, where respondents are visited repeatedly, sometimes by the same interviewer, sometimes by a different interviewer. Interviewer characteristics may affect the respondents' cooperation, which is analyzed with a multilevel model, with respondents nested within interviewers (Hox, de Leeuw, & Kreft, 1991). In longitudinal studies, the previous interviewer may also be relevant, and as a result, we have a cross-classified structure, with respondents nested within the cross-classification of the current and the previous interviewer. Examples of multilevel cross-classified analyses in panel interview studies are the studies by Pickery and Loosveldt (1998), O'Muircheartaigh and Campanelli (1999), and Pickery, Loosveldt, and Carton (2001). Cross-classified multilevel models have also been applied to sociometric choice data, where members of groups both give popularity ratings to and receive ratings from other group members (Van Duijn, van Busschbach, & Snijders, 1999). For other examples see Raudenbush (1993b) and Rasbash and Goldstein (1994).

Considering multilevel analysis of longitudinal data, the usual approach is to specify this as a hierarchical structure with measurement occasions nested within individual subjects. However, if all subjects are measured in the same time period (for example yearly periods all starting in the same year), it also makes sense to view the data as a cross-classification of measurement occasions with individual subjects, with factors 'subjects' and 'measurement occasions' both treated as random. Such a specification still allows missing data, in the form of panel dropout or occasionally missing a measurement occasion. Since the usual representation of measurement occasions nested within subjects is both efficient and corresponds to other representations as the

MANOVA model or latent curve modeling (see Chapter 16), this representation is usually preferred.

## 9.1 EXAMPLE OF CROSS-CLASSIFIED DATA: PUPILS NESTED WITHIN (PRIMARY AND SECONDARY SCHOOLS)

Assume that we have data from 1000 pupils who have attended 100 different primary schools, and subsequently went on to 30 secondary schools. Similar to the situation where we have pupils within schools and neighborhoods, we have a cross-classified structure. Pupils are nested within primary and within secondary schools, with primary and secondary schools crossed. In other words: pupils are nested within the cross-classification of primary and secondary schools. Goldstein (1994) introduces a formal description of these models, which will be followed here. In our example, we have a response variable *achievement* which is measured in secondary school. We have two explanatory variables at the pupil level: *pupil gender* (0 = male, 1 = female) and a six-point scale for pupil socioeconomic status, *pupil ses*. We have at the school level a dichotomous variable that indicates if the school is public (denom = 0) or denominational (denom = 1). Since we have both primary and secondary schools, we have two such variables (named *pdenom* for the primary school and *sdenom* for the secondary school).

At the pupil level, we can write an intercept-only model as:

$$Y_{i(jk)} = \beta_{0(jk)} + e_{i(jk)}, \tag{9.1}$$

where the achievement score of pupil $Y_{i(jk)}$ of pupil $i$ within the cross-classification of primary school $j$ and secondary school $k$ is modeled by the intercept (the overall mean) $\beta_{0(jk)}$ and a residual error term $e_{i(jk)}$. The subscripts ($jk$) are written within parentheses to indicate that they are conceptually at the same level: the ($jk$)th primary school/secondary school combination in the cross-classification of primary and secondary schools.

The subscripts ($jk$) indicate that we assume that the intercept $\beta_{0(jk)}$ varies independently across both primary and secondary schools. Thus, we can model the intercept using the second-level equation:

$$\beta_{0(jk)} = \gamma_{00} + u_{0j} + v_{0k} \tag{9.2}$$

In equation 9.2, $u_{0j}$ is the residual error term for the primary schools, and $v_{0k}$ is the residual error term for the secondary schools. After substitution, this produces the intercept-only model:

$$Y_{i(jk)} = \gamma_{00} + u_{0j} + v_{0k} + e_{i(jk)} \tag{9.3}$$

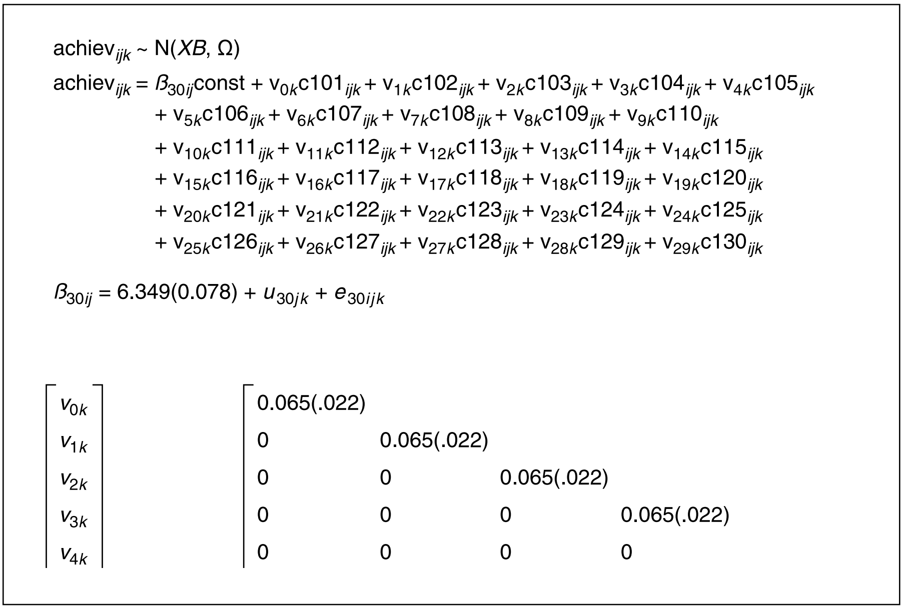
where the outcome variable is modeled with an overall intercept $\gamma_{00}$, with a residual error term $u_{0j}$ for primary school $j$ and a residual error term $v_{0k}$ for secondary school $k$, and the individual residual error term $e_{i(jk)}$ for pupil $i$ in the cross-classification of primary school $j$ and secondary school $k$.

Individual-level explanatory variables can be added to the equation, and their regression slopes may be allowed to vary across primary and/or secondary schools. School-level variables can also be added, and used to explain variation in the slopes of individual-level variables across schools, in a manner similar to ordinary multilevel regression models. The details of setting up a cross-classified model depend strongly on the software: some software (e.g., MLwiN) requires that users set up the model by hand; other software (e.g., HLM and SPSS) includes cross-classified models as part of the standard setup. Some comments on software use are given in the last section of this chapter.

Cross-classified models can be set up in all multilvel analysis software that allows equality constraints on the variance components. Raudenbush (1993b) and Rasbash and Goldstein (1994) show how a cross-classified model can be formulated and estimated as a hierarchical model. The example below uses MLwiN. Similar analyses can be carried out in other software, but these tend to hide the complications of the analysis setup. Ignoring the secondary school level, we can specify the individual and primary school level as usual, with individuals at the first level and primary schools at the second level. To create a place to specify the crossed effects of the secondary school level, we introduce a third 'dummy' level, with only one unit that covers the entire data set. At the pupil level, we specify a full set of dummy variables to indicate all of the secondary schools. In our example, we need 30 such dummy variables to indicate our 30 secondary schools. The fixed regression coefficients of these dummies are excluded from the model, but their slopes are allowed to vary at the third, the 'dummy', level. Subsequently, the covariances between these dummies are all constrained to be zero, and the variances are all constrained to be equal. Thus, in the end we estimate one variance component for the secondary schools, and by putting the secondary schools in a separate level we assure that there are no covariances between residuals for the primary and the secondary schools.

Although the analysis is set up using three separate levels, it should be clear that conceptually we have two levels, with primary and secondary schools crossed at the second level. The third level is just a computational device to allow estimation using standard multilevel software. For that reason, the third level is sometimes referred to as a 'dummy' level, and not the 'secondary school' level. There is no implication that the primary schools are nested within the secondary schools.

As an example, Figure 9.1 shows part of the equations window of MLwiN for a cross-classified analysis of our example data. The equations in Figure 9.1 are the MLwiN-equivalent of the cross-classified intercept-only model in equation 9.3.



$$\text{achiev}_{ijk} \sim \text{N}(XB, \Omega)$$

$$\begin{aligned}\text{achiev}_{ijk} = \beta_{30ij}\text{const} &+ \text{v}_{0k}\text{c101}_{ijk} + \text{v}_{1k}\text{c102}_{ijk} + \text{v}_{2k}\text{c103}_{ijk} + \text{v}_{3k}\text{c104}_{ijk} + \text{v}_{4k}\text{c105}_{ijk}\\ &+ \text{v}_{5k}\text{c106}_{ijk} + \text{v}_{6k}\text{c107}_{ijk} + \text{v}_{7k}\text{c108}_{ijk} + \text{v}_{8k}\text{c109}_{ijk} + \text{v}_{9k}\text{c110}_{ijk}\\ &+ \text{v}_{10k}\text{c111}_{ijk} + \text{v}_{11k}\text{c112}_{ijk} + \text{v}_{12k}\text{c113}_{ijk} + \text{v}_{13k}\text{c114}_{ijk} + \text{v}_{14k}\text{c115}_{ijk}\\ &+ \text{v}_{15k}\text{c116}_{ijk} + \text{v}_{16k}\text{c117}_{ijk} + \text{v}_{17k}\text{c118}_{ijk} + \text{v}_{18k}\text{c119}_{ijk} + \text{v}_{19k}\text{c120}_{ijk}\\ &+ \text{v}_{20k}\text{c121}_{ijk} + \text{v}_{21k}\text{c122}_{ijk} + \text{v}_{22k}\text{c123}_{ijk} + \text{v}_{23k}\text{c124}_{ijk} + \text{v}_{24k}\text{c125}_{ijk}\\ &+ \text{v}_{25k}\text{c126}_{ijk} + \text{v}_{26k}\text{c127}_{ijk} + \text{v}_{27k}\text{c128}_{ijk} + \text{v}_{28k}\text{c129}_{ijk} + \text{v}_{29k}\text{c130}_{ijk}\end{aligned}$$

$$\beta_{30ij} = 6.349(0.078) + u_{30jk} + e_{30ijk}$$

$$\begin{bmatrix} v_{0k} \\ v_{1k} \\ v_{2k} \\ v_{3k} \\ v_{4k} \end{bmatrix} \quad \begin{bmatrix} 0.065(.022) & & & \\ 0 & 0.065(.022) & & \\ 0 & 0 & 0.065(.022) & \\ 0 & 0 & 0 & 0.065(.022) \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

*Figure 9.1* (Part of) MLwiN specification of cross-classified, intercept-only model.

The first line in Figure 9.1 specifies a normal distribution for the outcome variable *achievement*. The second line and its continuation specifies the fixed part of the multilevel regression model. The model is the intercept-only model, and we see a term $\beta_{ij}$const for the intercept (for simplicity, we ignore the subscript 30 for the sequence number of the regression coefficient that corresponds to the intercept). The line $\beta_{ij}$ = 6.349 (0.078) + $u_{jk}$ + $e_{ijk}$ (again ignoring in our notation the subscript 30) gives the estimate for the intercept and its standard error, and shows that we assume two residual error terms: $e_{ijk}$ at the individual level, and $u_{jj}$ at the primary school level. What is not immediately clear in the equations window is that the 30 dummies for the secondary schools, labeled c101 to c130, are not in the fixed part of the regression model. The regression coefficients for v101 to v130 are not estimated, which is the same as stating that they are constrained to be zero. What is estimated is the variance of the residual error term $v_k$ at the secondary school level. We find this in the covariance matrix at the third, the dummy, level. Part of that 30 × 30 matrix is visible in Figure 9.1, and we can see that the off-diagonal elements are all constrained to be zero, and that the variance estimates on the diagonal are all constrained to be equal. The common variance estimate for the 30 dummies of 0.065 (s.e. .022) is the estimate for the variance of $v_{0k}$ at the secondary school level.

The results of a series of models on the cross-classified data set are presented in Table 9.1 in a more conventional form. The first column in Table 9.1 presents the results for the intercept-only model. Since cross-classified models usually contain more than two levels, which are not all unambiguously nested, the table does not use the usual sigma terms ($\sigma_e^2$, $\sigma_{u0}^2$, and so on) for the variance components, but names that correspond to the proper variable and level. Therefore, the term $\sigma^2_{\text{int/pupil}}$ corresponds to the usual lowest-level error term for the intercept $\sigma_e^2$ in the model equation and $\sigma^2_{\text{int/primary}}$ to the usual second-level error term for the intercept $\sigma_{u0}^2$. The term $\sigma^2_{\text{ses/primary}}$ refers to the second-level variance for the *ses* slope. When a model contains many variance components, indicating these in the results tables by proper names instead of symbols often makes interpretation easier.

*Table 9.1* Cross-classified model achievement in primary and secondary schools

| Model | Intercept-only<br>Coeff. (s.e.) | + pupil vars<br>Coeff. (s.e.) | + school vars<br>Coeff. (s.e.) | + *ses* random<br>Coeff. (s.e.) |
|---|---|---|---|---|
| **Fixed part** | | | | |
| Intercept | 6.35 (.08) | 5.76 (.11) | 5.52 (.19) | 5.53 (.14) |
| Pupil gender | | 0.26 (.05) | 0.26 (.05) | 0.25 (.05) |
| Pupil ses | | 0.11 (.02) | 0.11 (.02) | 0.11 (.02) |
| Primary denom | | | 0.20 (.12) | 0.20 (.12) |
| Secondary denom | | | 0.18 (.10) | 0.17 (.09) |
| **Random part** | | | | |
| $\sigma^2_{\text{int/pupil}}$ | 0.51 (.02) | 0.47 (.02) | 0.47 (.02) | 0.46 (.02) |
| $\sigma^2_{\text{int/primary}}$ | 0.17 (.04) | 0.17 (.04) | 0.16 (.04) | 0.14 (.08) |
| $\sigma^2_{\text{int/secondary}}$ | 0.07 (.02) | 0.06 (.02) | 0.06 (.02) | 0.05 (.02) |
| $\sigma^2_{\text{ses/primary}}$ | | | | 0.008 (.004) |
| Deviance | 2317.8 | 2243.5 | 2237.5 | 2224.5 |
| AIC | 2325.8 | 2255.5 | 2253.5 | 2244.5 |

Since the levels of the primary and secondary school are independent, we can add the estimated variances in the intercept-only model (the first column in Table 9.1) for a total variance of 0.75. The intraclass correlation for the primary school level is 0.17/0.75 = .23, and the intraclass correlation for the secondary school level is 0.07/ 0.75 = .09. So, 23% of the total variance is accounted for by the primary schools, and 9% by the secondary schools. Taken together, the schools account for (0.17 + 0.07)/0.75 = 0.32 of the total variance.

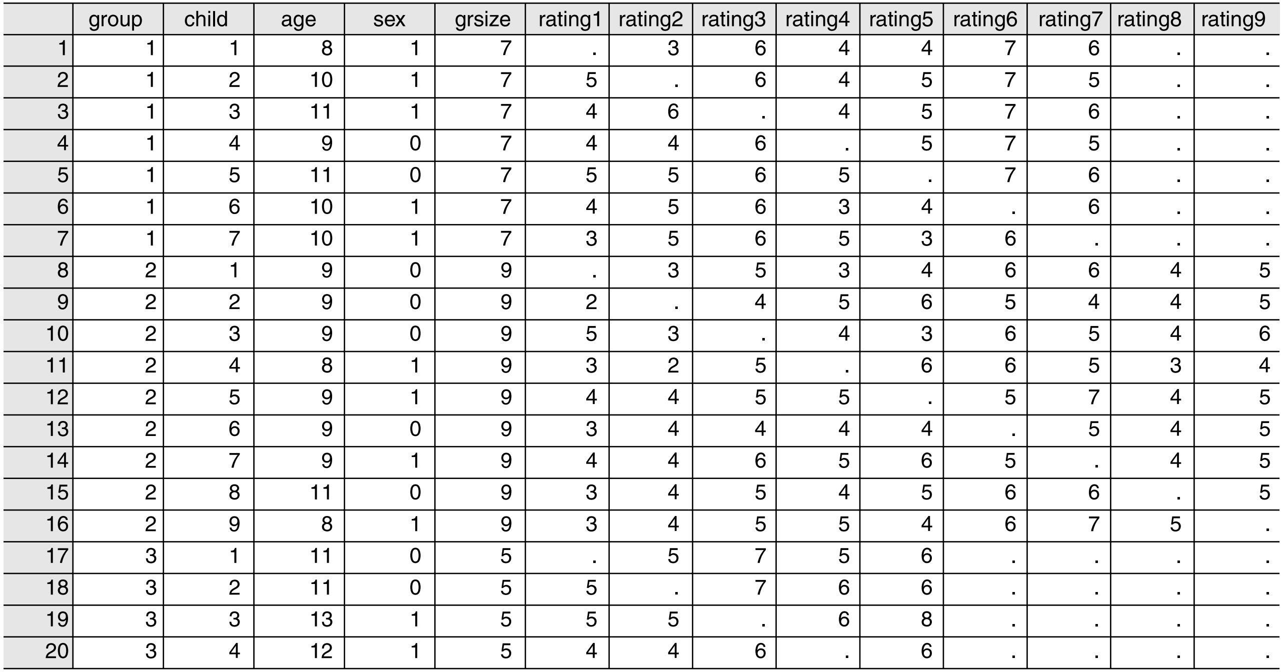
The pupil-level variables pupil gender and pupil ses have a significant contribution. The effect of either the primary or the secondary school being denominational is of borderline significance. The difference between the deviances of the second and the third model is 5.85, with two more parameters estimated in the third model. A chi-square test for this difference is also of borderline significance ($\chi^2 = 5.85$, $df = 2$, $p = .054$). The AIC indicates that we should prefer model 3. The conclusion is that, although there is apparently an effect of both school levels, the denomination of the school does not explain the school-level variance very well. The fourth column in Table 9.1 shows the estimates for the model where we allow for variation in the slope for pupil ses across primary schools. This is indicated by the term $\sigma^2_{\text{ses/primary}}$. There is a small but significant slope variance for *ses* at the primary school level. The slope variance at the secondary school level (not shown in the table) is negligible.

A drawback of formulating cross-classified models in this way is that the covariance matrix at the third level becomes very large when there are many units on that level. This may lead to computational problems when the number of units on the dummy level increases. For that reason, it is advantageous to use the dummy level for the classification with the fewest number of units.

## 9.2 EXAMPLE OF CROSS-CLASSIFIED DATA: (SOCIOMETRIC RATINGS) IN SMALL GROUPS

In the previous example, the cross-classification is at the higher levels, with pupils nested within the cross-classification of primary and secondary schools. The cross-classification can also be at lower levels. An example is the following model for sociometric ratings. Sociometric ratings can be collected by asking all members of a group to rate all other members, typically on a seven- or nine-point scale that indicates how much they would like to share some activity with the rated person. Figure 9.2 presents an example of a sociometric rating data set for three small groups, as it would look in standard statistical software.

In Figure 9.2, we see the sociometric ratings for a group of seven and a group of nine children, and part of the data of a third group of five children. High numbers indicate a positive rating. One way to collect such data is to give each child a questionnaire with a list of names for all children, and ask them to write their rating after each name. Therefore, each row in the table in Figure 9.2 consists of the sociometric ratings given by a specific child. The columns (variables) labeled *rating1*, *rating2*, . . . , *rating9* are the ratings received by child number 1, 2, . . . , 9. Figure 9.2 makes clear that network data, of which these sociometric ratings are an example, have a complicated structure that does not fit well in the rectangular data matrix assumed by most statistical software. The groups do not have the same size, so *rating6* to *rating9* have all



| | group | child | age | sex | grsize | rating1 | rating2 | rating3 | rating4 | rating5 | rating6 | rating7 | rating8 | rating9 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 8 | 1 | 7 | . | 3 | 6 | 4 | 4 | 7 | 6 | . | . |
| 2 | 1 | 2 | 10 | 1 | 7 | 5 | . | 6 | 4 | 5 | 7 | 5 | . | . |
| 3 | 1 | 3 | 11 | 1 | 7 | 4 | 6 | . | 4 | 5 | 7 | 6 | . | . |
| 4 | 1 | 4 | 9 | 0 | 7 | 4 | 4 | 6 | . | 5 | 7 | 5 | . | . |
| 5 | 1 | 5 | 11 | 0 | 7 | 5 | 5 | 6 | 5 | . | 7 | 6 | . | . |
| 6 | 1 | 6 | 10 | 1 | 7 | 4 | 5 | 6 | 3 | 4 | . | 6 | . | . |
| 7 | 1 | 7 | 10 | 1 | 7 | 3 | 5 | 6 | 5 | 3 | 6 | . | . | . |
| 8 | 2 | 1 | 9 | 0 | 9 | . | 3 | 5 | 3 | 4 | 6 | 6 | 4 | 5 |
| 9 | 2 | 2 | 9 | 0 | 9 | 2 | . | 4 | 5 | 6 | 5 | 4 | 4 | 5 |
| 10 | 2 | 3 | 9 | 0 | 9 | 5 | 3 | . | 4 | 3 | 6 | 5 | 4 | 6 |
| 11 | 2 | 4 | 8 | 1 | 9 | 3 | 2 | 5 | . | 6 | 6 | 5 | 3 | 4 |
| 12 | 2 | 5 | 9 | 1 | 9 | 4 | 4 | 5 | 5 | . | 5 | 7 | 4 | 5 |
| 13 | 2 | 6 | 9 | 0 | 9 | 3 | 4 | 4 | 4 | 4 | . | 5 | 4 | 5 |
| 14 | 2 | 7 | 9 | 1 | 9 | 4 | 4 | 6 | 5 | 6 | 5 | . | 4 | 5 |
| 15 | 2 | 8 | 11 | 0 | 9 | 3 | 4 | 5 | 4 | 5 | 6 | 6 | . | 5 |
| 16 | 2 | 9 | 8 | 1 | 9 | 3 | 4 | 5 | 5 | 4 | 6 | 7 | 5 | . |
| 17 | 3 | 1 | 11 | 0 | 5 | . | 5 | 7 | 5 | 6 | . | . | . | . |
| 18 | 3 | 2 | 11 | 0 | 5 | 5 | . | 7 | 6 | 6 | . | . | . | . |
| 19 | 3 | 3 | 13 | 1 | 5 | 5 | 5 | . | 6 | 8 | . | . | . | . |
| 20 | 3 | 4 | 12 | 1 | 5 | 4 | 4 | 6 | . | 6 | . | . | . | . |

*Figure 9.2* Sociometric rating data for three small groups.

missing values for group 3, which has only five children. The children do not rate themselves, so these ratings also have missing values in the data matrix. The data also include the pupil characteristics *age* and *gender* and the group characteristic *group size*.

Special models have been proposed for network data (for an extensive introduction see Wasserman & Faust, 1994), and specialized software is available for the analysis of network data. Van Duijn, Snijders, and Zijlstra (2004) show that one can also use multilevel regression models to analyze network data. In the example of the sociometric ratings, we would view the ratings as an outcome variable that is nested within the cross-classification of the *senders* and the *receivers* of sociometric ratings. At the lowest level we have the separate ratings that belong to specific sender–receiver pairs. This is nested within the cross-classification of senders and receivers at the second level, which in turn can be nested, for example within a sample of groups.

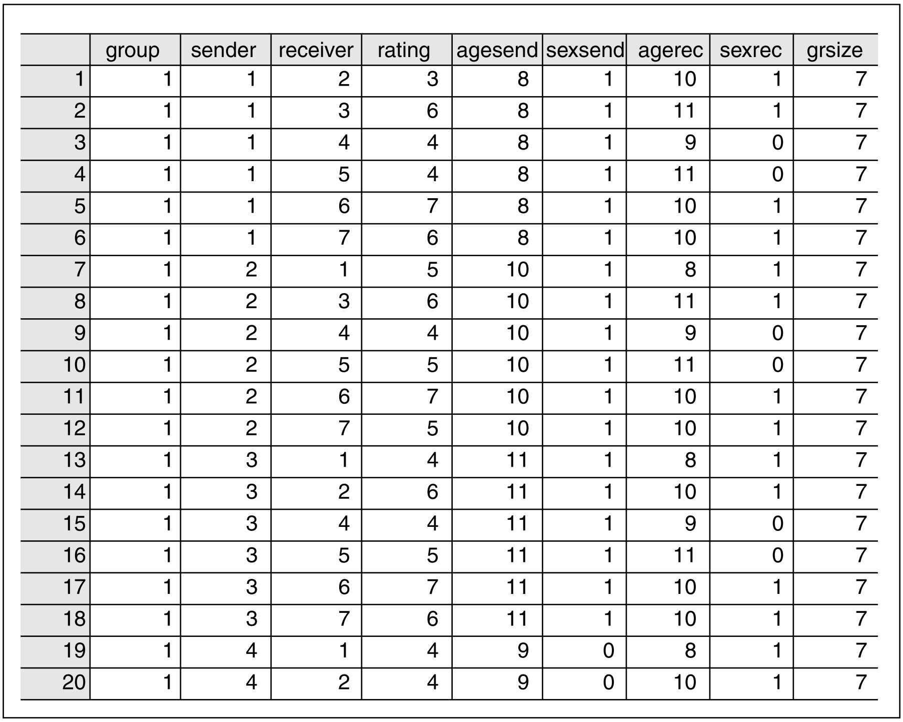
To analyze sociometric choice data using multilevel techniques, the data must be arranged in a different format than the one shown in Figure 9.2. In the new data file, the individual rows must refer to the separate ratings, with associated child identification codes to identify the sender and receiver in that particular rating, and the variables that characterize the sender and receiver of information. Such a data set looks like the one depicted in Figure 9.3. This figure illustrates clearly how the distinct ratings are nested below both senders and receivers, who in turn are nested below the sociometric groups.

As the data set in Figure 9.3 illustrates, the cross-classification here is at the second level. We have ratings of sender and receiver pairs, which form a cross-classification nested within the groups. At the lowest level are the separate ratings. At the second level, we have the explanatory variables *age* and *gender* for the senders and receivers of ratings, and at the group level, we have the group characteristic *group size*.

The data in Figures 9.2 and 9.3 are part of a data set that contains sociometric data from 20 groups of children, with group sizes varying between four and eleven. At the lowest level, we can write the intercept-only model as follows:

$$Y_{i(jk)l} = \beta_{0(jk)l} + e_{i(jk)l} \tag{9.4}$$

In equation 9.4, rating $i$ of sender $j$ and receiver $k$ is modeled by an intercept $\beta_{0(jk)l}$. At the lowest level, the ratings level, we have residual random errors $e_{i(jk)l}$, which indicates that we do not assume that all variation between ratings can be explained by differences between senders and receivers. These residual errors could be the result of random measurement errors, but they could also reflect unmodeled interactions between senders and receivers. The cross-classification of senders and receivers is nested within the groups, indicated by $l$. Again, parentheses are used to indicate a cross-classification



| | group | sender | receiver | rating | agesend | sexsend | agerec | sexrec | grsize |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 2 | 3 | 8 | 1 | 10 | 1 | 7 |
| 2 | 1 | 1 | 3 | 6 | 8 | 1 | 11 | 1 | 7 |
| 3 | 1 | 1 | 4 | 4 | 8 | 1 | 9 | 0 | 7 |
| 4 | 1 | 1 | 5 | 4 | 8 | 1 | 11 | 0 | 7 |
| 5 | 1 | 1 | 6 | 7 | 8 | 1 | 10 | 1 | 7 |
| 6 | 1 | 1 | 7 | 6 | 8 | 1 | 10 | 1 | 7 |
| 7 | 1 | 2 | 1 | 5 | 10 | 1 | 8 | 1 | 7 |
| 8 | 1 | 2 | 3 | 6 | 10 | 1 | 11 | 1 | 7 |
| 9 | 1 | 2 | 4 | 4 | 10 | 1 | 9 | 0 | 7 |
| 10 | 1 | 2 | 5 | 5 | 10 | 1 | 11 | 0 | 7 |
| 11 | 1 | 2 | 6 | 7 | 10 | 1 | 10 | 1 | 7 |
| 12 | 1 | 2 | 7 | 5 | 10 | 1 | 10 | 1 | 7 |
| 13 | 1 | 3 | 1 | 4 | 11 | 1 | 8 | 1 | 7 |
| 14 | 1 | 3 | 2 | 6 | 11 | 1 | 10 | 1 | 7 |
| 15 | 1 | 3 | 4 | 4 | 11 | 1 | 9 | 0 | 7 |
| 16 | 1 | 3 | 5 | 5 | 11 | 1 | 11 | 0 | 7 |
| 17 | 1 | 3 | 6 | 7 | 11 | 1 | 10 | 1 | 7 |
| 18 | 1 | 3 | 7 | 6 | 11 | 1 | 10 | 1 | 7 |
| 19 | 1 | 4 | 1 | 4 | 9 | 0 | 8 | 1 | 7 |
| 20 | 1 | 4 | 2 | 4 | 9 | 0 | 10 | 1 | 7 |

*Figure 9.3* Sociometric data rearranged for multilevel analysis, first four senders.

of factors that are conceptually at the same level: the $(jk)$th sender/receiver combination, which is nested within group $l$.

Note that we use subscripts on the intercept term $\beta_0$ to indicate that we assume that the intercept varies across both senders and receivers. Models involving cross-classified levels tend to have many distinct levels, and the practice of assigning a different Greek letter to regression coefficients at each level leads in such cases to a confusing array of Greek letters. In this chapter, the Greek letter $\beta$ is used for regression coefficients that are assumed to vary across some level(s), with subscripts indicating these levels, and the Greek letter $\gamma$ is used to denote fixed regression coefficients. So, the subscripts $j$, $k$, and $l$ on the regression coefficient $\beta_{0(jk)l}$ indicate that we assume that the intercept $\beta_{0(jk)l}$ varies across the cross-classification of senders and receivers nested within groups. Thus, we can model this intercept variance using the second-level equation:

$$\beta_{0(jk)l} = \beta_{0l} + u_{0j} + v_{0kl}. \tag{9.5}$$

The subscript $l$ on the regression coefficient $\beta_{0l}$ indicates that we assume that the intercept $\beta_{0l}$ varies across groups. We can further model the intercept variance using the third-level equation:

$$\beta_{0l} = \gamma_{00} + f_{0l}. \tag{9.6}$$

After substitution, this produces:

$$Y_{i(jk)l} = \gamma_{00} + f_{0l} + u_{0jl} + v_{0kl} + e_{i(jk)l}, \tag{9.7}$$

where the outcome variable is modeled with an overall intercept $\gamma_{00}$, together with a residual error term $f_l$ for group $l$, the individual-level residual error terms $u_{jl}$ for sender $j$ in group $l$ and $v_{kl}$ for receiver $k$ in group $l$, and the measurement-level error term $e_{i(jk)l}$.

Since there are as many senders as there are receivers, it is immaterial whether we use the dummies to indicate the senders or the receivers. Here, the crossed effects of the receiver level are incorporated using dummies. At the lowest level, the ratings, we specify dummy variables that indicate the receivers. In our example data, the largest group consists of 11 pupils, so we need 11 dummy variables. The fixed coefficients of these dummies are excluded from the model, but their slopes are allowed to vary. Since the cross-classification is nested within the sociometric groups, the slopes of the dummy variables are set to vary at a third group level defined by the group identification variable. In addition, the covariances between the receiver dummies are constrained to be zero, and the variances are constrained to be equal. Thus, we estimate one variance component for the receivers, and by putting the variance term(s) for the receivers on a separate level we assure that there are no covariances between the residuals for the sender and the receiver level. Both sender and receiver characteristics like *age* and *gender* and group characteristics like *group size* can be added to the model as predictors, and child characteristics may be allowed to have random slopes at the group level. The analysis proceeds along exactly the same lines as outlined for the cross-classification of primary and secondary schools.

Since the third 'group' level is already used to specify the random variation for the receiver dummies, we must make sure that the intercept and possible slope variation at the 'real' group level are not correlated with the dummies. This can be accomplished by adding the appropriate constraints to the model. When the software supports more than three levels (e.g., MLwiN), the same result can be accomplished more conveniently by adding a fourth level to the model; also for the groups, which is used for the random part at the real group level. Once more, it should be emphasized that, although the analysis is then set up using four separate levels, conceptually we have three levels, with senders and receivers crossed at the second level. The fourth level, the duplicated group level, is just a computational device to allow simple estimation of these very complex models using standard multilevel software.

To analyze our sociometric data, we decide to define the senders as an ordinary level in the standard manner, and the receivers at the third, the 'dummy', group level, using dummies. The first model is the intercept-only model of 9.7 written with variable names rather than symbols.

$$rating_{i(jk)l} = \gamma_{00} + f_{0l} + u_{0jl} + v_{0kl} + e_{i(jk)l} \tag{9.8}$$

This produces an estimate for the overall intercept, and the variance components $\sigma_e^2$ for the variance of the ratings, $\sigma_{u_0}^2$ and $\sigma_{v_0}^2$ for the senders and the receivers, plus $\sigma_{f_0}^2$ for the variance at the group level. The estimates are in the first column of Table 9.2.

*Table 9.2* Results for cross-classified model of sociometric ratings in groups

| Model | Intercept-only | + all fixed | + sender gender random | + interaction gender/gender |
|---|---|---|---|---|
| | Coeff. (s.e.) | Coeff. (s.e.) | Coeff. (s.e.) | Coeff. (s.e.) |
| **Fixed part** | | | | |
| Intercept | 5.03 (.22) | 1.56 (1.17) | 1.00 (1.00) | 1.00 (1.00) |
| Sender age | | 0.23 (.03) | 0.22 (.03) | 0.22 (.03) |
| Sender gender | | –0.16 (.07) | –0.12 (.13) | –0.37 (.14) |
| Receiver age | | 0.21 (.06) | 0.21 (.06) | 0.22 (.06) |
| Receiver gender | | 0.74 (.13) | 0.73 (.13) | 0.49 (.13) |
| Group size | | –0.17 (.10) | –0.08 (.07) | –0.08 (.07) |
| Interaction gender/ gender | | | | 0.51 (.09) |
| **Random part** | | | | |
| $\sigma^2_{\text{int/ratings}}$ | 0.41 (.02) | 0.42 (.02) | 0.42 (.02) | 0.40 (.02) |
| $\sigma^2_{\text{int/senders}}$ | 0.15 (.03) | 0.09 (.01) | 0.02 (.01) | 0.02 (.01) |
| $\sigma^2_{\text{int/receivers}}$ | 0.65 (.09) | 0.49 (.07) | 0.49 (.07) | 0.48 (.07) |
| $\sigma^2_{\text{int/groups}}$ | 0.84 (.30) | 0.42 (.16) | 0.23 (.10) | 0.23 (.10) |
| $\sigma^2_{\text{send.gend/groups}}$ | | | 0.28 (.11) | 0.30 (.11) |
| $\sigma^2_{\text{send.gend-int/groups}}$ | | | 0.17 (.08) | 0.18 (.08) |
| Deviance | 2772.4 | 2677.0 | 2613.7 | 2580.3 |
| AIC | 2782.4 | 2697.3 | 2637.7 | 2606.2 |

For the sake of readability, the variance components in the random part are indicated by proper variable and level names instead of the usual symbols. From the intercept-only model in the first column, it appears that 20% of the total variance is at the lowest (ratings) level, only 7% of the total variance is variance between the senders, 32% of the total variance is variance between the receivers, and 41% is variance between groups. Apparently, there are strong group effects.

The model in the second column of Table 9.2 adds all available explanatory variables as fixed predictors. Using abbreviated variable names, it can be written as:

$$\begin{aligned} rating_{i(jk)l} = \gamma_{00} + \gamma_{10}\, send.age_{jl} + \gamma_{20}\, send.gender_{jl} + \gamma_{30} rec.age_{kl} + \gamma_{40} \\ rec.gender_{kl} + \gamma_{01}\, groupsize_l + f_{0l} + u_{0jl} + v_{0kl} + e_{i(jk)l} \end{aligned} \tag{9.9}$$

From the regression coefficients estimated in the second column of Table 9.2 we conclude that age of sender and receiver have a small positive effect on the ratings. In larger groups, the ratings are a bit lower, but this effect is not significant. The effect of the children's gender is more complicated. *Gender* is coded 0 for male, 1 for female, and Table 9.2 shows that female senders give lower ratings, while female receivers get higher ratings.

Only one explanatory variable (*send.gender*) has a significant slope variation at the group level. This model can be written as:

$$\begin{aligned} rating_{i(jk)l} = \gamma_{00} + \gamma_{10}\, send.age_{jl} + \gamma_{20}\, send.gender_{jl} + \gamma_{30} rec.age_{kl} + \gamma_{40} \\ rec.gender_{kl} + \gamma_{01}\, groupsize_l + f_{1l}\, send.gender_{jl} + f_{0l} + u_{0jl} + v_{0kl} + e_{i(jk)l} \end{aligned} \tag{9.10}$$

Apart from having two residual error terms at the second level, one for the senders and one for the receivers, equation 9.10 represents an ordinary multilevel regression model with one varying slope at the group level, and no cross-level interactions. The estimates for this model are in the third column of Table 9.2. The variance of the slopes for sender gender is substantial, which indicates that the effect of the gender of the senders differs considerably across groups.

The only group-level variable available to explain the variation of the *sender gender* slopes is group size. However, if we enter that in the model by forming the cross-level interaction between the variables *sender gender* and *group size*, it turns out to be nonsignificant.

The different effects of sender and receiver gender suggest looking in more detail at the effect of the explanatory variable *gender*. The last model in Table 9.2 includes an interaction effect for sender gender and receiver gender. This interaction, which is an ordinary interaction and not a cross-level interaction, is indeed significant. To interpret this interaction, it is useful to construct a graph. Figure 9.4 shows the interaction. It is clear that, in addition to the direct effects of sender gender (girls give on average lower

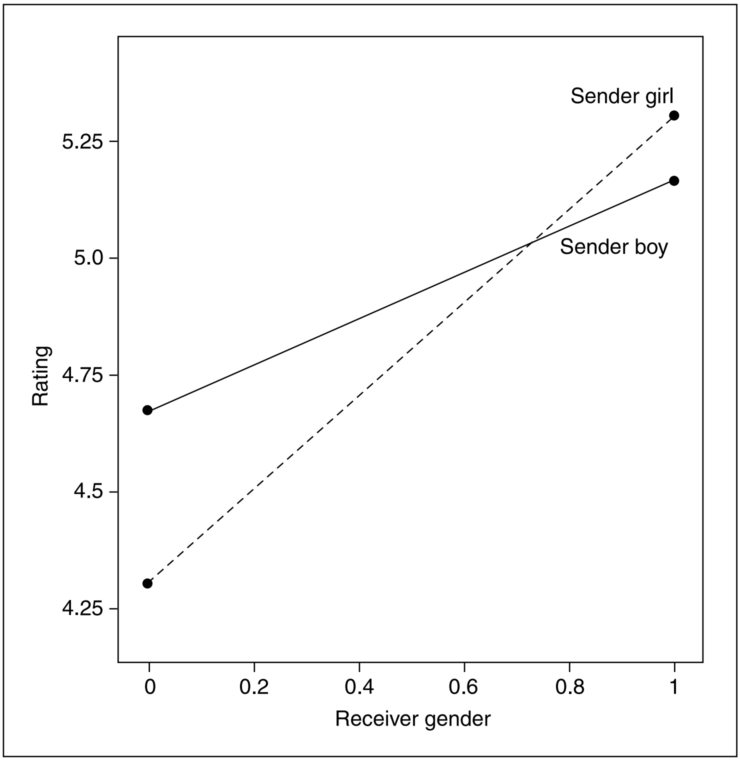


*Figure 9.4* Graph of interaction between sender gender and receiver gender.

ratings than boys) and receiver gender (girls receive on average higher ratings than boys), there is an interaction effect: both boys and girls give higher ratings to other children from their own gender.

Snijders and Bosker (1999, Chapter 11) discuss a number of extensions of multilevel models for sociometric data. For instance, it may be useful to insert a second level that defines the *dyads*, the sender–receiver pairs. For each pair there are two ratings. The amount of variance at the dyad level indicates the degree of reciprocity in the ratings. For other extensions to the multilevel analysis of network data see Snijders and Kenny (1999) and Van Duijn et al. (1999). The analysis of so-called ego-centered network data, which have a simple nesting structure, is discussed by Snijders, Spreen, and Zwaagstra (1994), Spreen and Zwaagstra (1994), and by Kef, Habekothé, and Hox (2000). Models for dyadic data are discussed extensively by Kenny, Kashy, and Cook (2006).

## 9.3 STATISTICAL AND COMPUTATIONAL ISSUES

It is clear that for cross-classified models we need software that allows creating a highest level with just one unit, and imposing zero and equality constraints on the highest-level variances and covariances. In addition, we are going to have large and complicated models. For instance, in our example with the cross-classification of primary and secondary schools, we are modeling a single variance component for the secondary school level using 30 dummy variables. For the random variance of the intercept at the secondary school level, we need therefore a covariance matrix for the random intercept of 30 × 30 = 900 elements. This covariance matrix must have equality constraints on the 30 variances on the diagonal, and constrain the 435 covariances to be all zero. If we add a pupil-level variable, for instance *pupil ses*, and allow its slope to vary across secondary schools, the model becomes even more complicated. To create the secondary school-level dummies for this variance component, we must create 30 new *pupil ses* dummies that are the product of the 30 secondary school dummies and the value of *pupil ses*. These dummies are then added to the model, and allowed to vary at the highest, the 'dummy', level. We now have a covariance matrix at the secondary school level of 60 × 60 = 3600 elements. We must constrain the 30 variances of the intercept dummies to be equal, and their 435 covariances to be zero. In addition, we must constrain the 30 variances of the slopes of the *ses* dummies to be equal, and their 435 covariances to be zero. In general, we assume that the covariance between the intercepts and slopes at a higher level is not zero, which means that we must estimate in our example a common value by imposing equality constraints on the 30 × 30 = 900 covariances between intercept and slope dummies. This is an intimidating prospect, and even more so if we had, say, 100 secondary schools.

It is clear that there are practical limits to what can be done in standard multilevel software. For the model, it does not matter whether we define a dummy level for the primary or the secondary schools. The way the cross-classified model is set up, they are completely exchangeable, and we should get the same estimates either way. In practice, if we specify the 100 primary schools as a dummy level, we need 100 dummy variables. It is clear that in general we would prefer to use the dummy level for the smallest classification, which in our example is the secondary school level. To specify the 30 secondary schools as a dummy level, we need only 30 dummy variables.

Given the complexity and size of cross-classified models, it is advisable to carry out preliminary analyses using standard multilevel models, leaving out each of the cross-classified factors in turn. This provides a rough indication of which variance components are important at which level. This is only a rough indication, because if the two classifications are correlated, leaving one out will lead to upwardly biased variance estimates for the other classification. Nevertheless, it will provide some indication of which classification has the largest variance components. Since specifying varying slopes on the dummy level is complicated and time consuming, it is best to use the standard specification for the classification with the largest number of random effects, and specify the other classification as a dummy level. In fact, if we carry out three-level analyses on our sociometric ratings data, with ratings at the lowest level, either sender or receiver at the second level, and groups at the third level, we find that sender-level variables appear to have significant variance components at the group level, while receiver-level variables appear to have only fixed effects. This preliminary result justifies our decision to use a regular level for the senders, and model the receivers using a dummy level.

Specifying cross-classified models is straightforward in the command mode of MLwiN because it contains a command SETX that carries out the tedious details. Other multilevel programs can be used, provided that they have facilities to impose constraints on the covariances and variances of the dummies. For instance, cross-classified models can be fitted in SAS and STATA using essentially the same approach as explained earlier in this chapter. Also, HLM (from version 6.0) and SPSS (from version 11.5) can directly specify cross-classified models. In HLM the process is straightforward. In SPSS cross-classified models can be set up in the syntax window. However, in SPSS the RANDOM command behaves differently depending on the coding of the lowest-level units. Unique identification numbers have different effects than restarting the count in each group. SPSS users are referred to the review by Leyland (2004). Although this direct setup of cross-classified models does not involve constructing a large set of dummy variables for one of the classifications, the issues are similar. So, HLM distinguishes between 'row' and 'column' classifications, and directs the user to define the identification variable with the smallest number of groups as the 'column' variable, to speed up computations. In SPSS, different specifications of the same model lead to the same estimates, but very different computation times (Leyland, 2004).

There is a second way to simplify a cross-classified model. Our 30 secondary schools could be crossed with the primary schools in such a way that we can distinguish, for instance, three groups of each 10 secondary schools that receive pupils from, say, 34, 33, and 33 of the primary schools, in such a way that this part of the structure is completely nested. That is, the first group of secondary schools takes only pupils from the first 34 primary schools, the second group takes only pupils from the following 33 primary schools, and the third group takes only pupils from the last 33 primary schools. This means that we can split the data set at the secondary school level into three separate groups. If we do this, we can simplify the model by specifying the third level as the secondary school-group level. We now need only 10 dummies for the secondary schools (the largest number of secondary schools in any of the groups so made), and the covariance matrix reduces to a size of 10 × 10 = 100 elements, instead of the 900 we had earlier. If we add the *ses* slope, we now need a covariance matrix of 20 ×

20 = 400 elements, instead of 3600. Finding such groupings dramatically reduces the size of the model. In fact, if there is no such separate grouping structure, but we can create it by deleting a few stray pupils from the data set, it may well be worth it. MLwiN contains a command to search for such groupings, but with other programs one has to tabulate the 100 × 30 cross-tabulation of the primary and secondary schools to search visually for separable groupings.

The examples all use 0/1 dummies to indicate group membership. If an individual is a member of more than one group, or if group membership is unknown for some individuals, it is possible to 'spread' the dummy value of 1.0 over several groups, for instance giving each of two possible groups a dummy value of 0.5. This kind of model, which is useful for multiple memberships or if group membership is fuzzy (see Blalock, 1990), is discussed by Hill and Goldstein (1998).

# 10

# Multivariate Multilevel Regression Models

Multivariate multilevel regression models are multilevel regression models that contain more than one response variable. As such, they are comparable to classical multivariate analysis of variance (MANOVA) models, where we also have several outcome measures. The reason for using a multivariate model is usually that the researchers have decided to use multiple measurements of one underlying construct, to achieve a better construct validity. A classic example is in medical research when diseases manifest themselves in a *syndrome* that leads to a pattern of related effects (Sammel, Lin, & Ryan, 1999). By using several outcome measures, researchers can obtain a better and more complete description of what is affected by changes in the predictor variables. Tabachnick and Fidell (2007) mention several advantages of using a multivariate approach instead of carrying out a series of univariate analyses. One advantage of multivariate analysis is that carrying out a series of univariate statistical tests inflates the type I error rate, which is controlled better in a multivariate analysis. A second advantage of multivariate analysis is that it often has more power. On each individual response measure, the differences may be small and insignificant, but for the total set of response measures, the joint effect may produce a significant effect (Stevens, 2009; Tabachnick & Fidell, 2007). However, the disadvantage of multivariate models is that they are more complicated, and that their interpretation is more ambiguous.

In multilevel analysis, using multiple outcome measures leads to some very powerful analysis tools. First, like in analysis of variance, using several response variables may lead to more power. Since multilevel analysis does not assume that all response measures are available for all individuals, it may be used as an alternative for MANOVA when there are missing values on some of the response variables. Most software for MANOVA cannot cope with missing data on the response variables, while for multilevel analysis this poses no special problem. Since the multilevel model is much more flexible than MANOVA, there are some additional advantages to multivariate multilevel modeling. For instance, since multivariate multilevel analysis combines multiple response variables in one model, it is possible to test the equality of their regression coefficients or variance components by imposing equality constraints. Also, the covariances between the dependent variables can be decomposed over the separate levels, which is one way to obtain the covariance matrices needed for multilevel factor analysis or structural equation modeling (see Chapters 14 and 15 for details). Finally, it is possible to construct multilevel measurement models, by including a set of questions that form a scale as multivariate responses in a multilevel model.

## 10.1 THE MULTIVARIATE MODEL

The multilevel regression model is inherently a univariate model. Even so, it can be used to analyze multiple outcome variables by placing these in a separate 'variables' level. In the multivariate multilevel regression model, the different measures are the lowest-level units. In most applications, the different measures would be the first level, the individuals the second level, and if there are groups, these form the third level. Therefore, if we have $p$ response variables, $Y_{hij}$ is the response on measure $h$ of individual $i$ in group $j$.

Ignoring possible missing responses, at the lowest level (the variable level) we have $p$ 'cases', which in fact are the $p$ response variables. Each case has a single response, which is the response of person $i$ to question $h$. One way to represent the different outcome variables would be to define $p - 1$ dummy variables that indicate the variables 2, . . ., $P$. In this scheme, variable 1 would be the base category, indicated by the value 0 for all dummy variables. However, this would give the first variable a special position. A much better way to indicate the multiple response variables is to leave out the intercept, and to define $p$ dummy variables, one for each response variable. Thus, we have $p$ dummy variables $d_{phij}$, defined for $p = 1, . . ., P$ by:

$$d_{phij} = \begin{cases} 1 & p = h \\ 0 & p \neq h \end{cases}. \tag{10.1}$$

To use these $p$ dummy variables in a model, we must exclude the usual intercept from the model. Hence, on the lowest level we have:

$$Y_{hij} = \pi_{1ij} d_{1ij} + \pi_{2ij} d_{2ij} + . . . + \pi_{pij} d_{pij}. \tag{10.2}$$

We use an extra level, the *dummy variable* level, to specify a multivariate model using software that is essentially developed for univariate analyses. There is no lowest-level error term in equation 10.2; the lowest level exists solely to define the multivariate response structure.[1] For the moment, we assume no explanatory variables, and we have the equivalent of the intercept-only model. Then, at the individual level (the second level in the multivariate model), we have:

[1] The symbol $\pi$ is used for the lowest-level regression coefficients, so we can continue to employ $\beta$ for the individual-level and $\gamma$ for the group-level regression coefficients.

$$\pi_{pij} = \beta_{pj} + u_{pij}. \tag{10.3}$$

At the group level (the third level in the multivariate model), we have:

$$\beta_{pj} = \gamma_p + u_{pj}. \tag{10.4}$$

By substitution we obtain:

$$\begin{aligned} Y_{hij} = {} & \gamma_1 d_{1ij} + \gamma_2 d_{2ij} + \ldots + \gamma_p d_{pij} \\ & + u_{1ij} d_{1ij} + u_{2ij} d_{2ij} + \ldots + u_{pij} d_{pij} + u_{1j} d_{1ij} + u_{2j} d_{2ij} + \ldots + u_{pj} d_{pij}. \end{aligned} \tag{10.5}$$

In the univariate intercept-only model, the fixed part contains only the intercept, which is the overall mean, and the random part contains two variances, which are the variances at the individual and the group level. In the equivalent multivariate model, the fixed part contains, in place of the intercept, the *P* regression coefficients for the dummy variables, which are the *P* overall means for the *P* outcome variables. The random part contains two covariance matrices, $\mathbf{\Omega_{ij}}$ and $\mathbf{\Omega_j}$, which contain the variances and the covariances of the regression slopes for the dummies at the individual and the group level. Since equation 10.5 is complicated, especially if we have many response variables, it is often expressed using sum notation:

$$Y_{hij} = \sum_{h=1}^{P} \gamma_h d_{hij} + \sum_{h=1}^{P} u_{hij} d_{hij} + \sum_{h=1}^{P} u_{hj} d_{hij}. \tag{10.6}$$

Just as in univariate modeling, explanatory variables at the individual or the group level can be added to the model. In general, we add an individual-level explanatory variable $X_{ij}$ or a group-level variable $Z_j$ to the model by multiplying it with all *p* dummy variables, and adding all *p* resulting interaction variables to the equation. Since the dummy variables are equal to zero whenever $p \neq h$, these terms disappear from the model. Thus there are *p* distinct contributions to the multilevel regression equation, each specific to one of the *p* response variables.

We can specify random slopes for the individual-level explanatory variables at the group level, and add cross-level interactions to explain random variation, completely analogous to adding explanatory variables and cross-level interactions to the univariate models discussed in Chapter 2. If we multiply each explanatory variable by all of the dummy variables, we allow each regression coefficient in the model to be different for each response variable. It would simplify the model considerably if we could impose an equality constraint across all response variables, assuming that the effects are equal for all response variables. There are two ways to accomplish this. For simplicity, let us assume that we have two response variables $Y_1$ and $Y_2$, only one explanatory variable $X$, and no group structure. Equation 10.2 now becomes:

$$Y_{hi} = \pi_{1i} d_{1i} + \pi_{2i} d_{2i}, \tag{10.7}$$

and equation 10.5 simplifies to:

$$Y_{hi} = \gamma_1 d_{1i} + \gamma_2 d_{2i} + u_{1i} d_{1i} + u_{2i} d_{2i}. \tag{10.8}$$

We add explanatory variable $X_i$ to the model, by multiplying it by each dummy variable. This produces:

$$Y_{hi} = \gamma_{01} d_{1i} + \gamma_{02} d_{2i} + \gamma_{11} d_{1i} X_i + \gamma_{12} d_{2i} X_i + u_{1i} d_{1i} + u_{2i} d_{2i}. \tag{10.9}$$

If we force the two regression coefficients for $Y_1$ and $Y_2$ to be equal by adding the constraint that $\gamma_{11} = \gamma_{12} = \gamma^*$, we get:

$$Y_{hi} = \gamma_{01} d_{1i} + \gamma_{02} d_{2i} + \gamma^* d_{1i} X_i + \gamma^* d_{2i} X_i + u_{1i} d_{1i} + u_{2i} d_{2i}, \tag{10.10}$$

which can also be written as:

$$Y_{hi} = \gamma_{01} d_{1i} + \gamma_{02} d_{2i} + \gamma^* [d_{1i} X_i + d_{2i} X_i] + u_{1i} d_{1i} + u_{2i} d_{2i}. \tag{10.11}$$

Since the two dummies that indicate the separate response variables are mutually exclusive, only one dummy variable will be equal to one for each specific response variable $Y_{hi}$, and the other is equal to zero. Therefore, equation 10.11 can also be written as:

$$Y_{hi} = \gamma_{01} d_{1i} + \gamma_{02} d_{2i} + \gamma^* X_i + u_{1i} d_{1i} + u_{2i} d_{2i}. \tag{10.12}$$

This makes it clear that imposing an equality constraint across all regression slopes for a specific explanatory variable is equal to adding this explanatory variable directly, without multiplying it by all the available dummies. This also implies that the model of equation 10.12 is nested within the model of equation 10.9. As a result, we can test whether simplifying model 10.9 to model 10.12 is justified, using the chi-square test on the deviances, with $p - 1$ degrees of freedom. The example given here involves changes to the fixed part, so we can use the deviance test only if we use FML estimation. If the explanatory variable $X$ has random slopes at the group level, a similar argument would apply to the random part of the model. Adding a random slope for one single explanatory variable $X$ to the model implies estimating one variance component. Adding a random slope to each of the explanatory variables constructed by multiplying $X$ by each of the dummies implies adding a $p \times p$ (co)variance matrix to the model. This adds $p(p - 1)/2$ parameter estimates to the model, and the degrees of freedom for the corresponding simultaneous chi-square difference test are $(p(p - 1)/2) - 1$.

## 10.2 EXAMPLE OF MULTIVARIATE MULTILEVEL ANALYSIS: MULTIPLE RESPONSE VARIABLES

Chapter 6 discusses an example that analyzes response rates on face-to-face, telephone, and mail surveys, as reported in 47 studies over a series of years (Hox & de Leeuw, 1994). In this example, there are two indicators of survey response. The first is the completion rate, which is the number of completed interviews divided by the total number of persons approached. The second is the response rate, which is the number of completed interviews divided by the total number of persons approached *minus* the number of persons that are considered ineligible (address incorrect, deceased). Some studies report the completion rate, some the response rate, and some both. The analysis reported in Chapter 6 analyzes response rates where available, and otherwise completion rates, with a dummy variable indicating when the completion rate is used. Since some studies report both the response rate and the completion rate, this approach is wasteful because it ignores part of the available information. Furthermore, it is an interesting question by itself: whether the response rate and completion rate behave similarly or differently over time. Using a multivariate model we can include all information, and carry out a multivariate meta-analysis to investigate the similarity between response rate and completion rate.

In Chapter 6 we have a two-level model, with data collection conditions (face-to-face, telephone, and mail) as the lowest level, and the 47 studies the second level. For the multivariate model, we specify the response as the lowest level. The conditions (face-to-face, telephone, and mail) are the second level and the studies are the third level. Since the response variable is a proportion, we use a generalized linear model with a logit link and a binomial error distribution (for details on multilevel generalized linear models see Chapter 6). Let $p_{hij}$ be the observed proportions of respondents on the response rate or completion rate in condition $i$ of study $j$. At the response indicator level, we have two explanatory variables, *comp* and *resp*, which are dummies that indicate whether the response is a completion rate or a response rate. The multivariate empty model can now be written as:

$$P_{hij} = \text{logistic}(\pi_{1ij} comp_{ij} + \pi_{2ij} resp_{ij}). \tag{10.13}$$

The empty model for these data is:

$$P_{hij} = \text{logistic}\begin{pmatrix}\gamma_{01}comp_{ij} + \gamma_{02}resp_{ij} \\ + u_{1ij}comp_{ij} + u_{2ij}resp_{ij} + u_{1j}comp_{ij} + u_{2j}resp_{ij}\end{pmatrix}. \tag{10.14}$$

The model of equation 10.14 provides us with estimates (on the logit scale) of the average completion rate and response rate, and the covariance matrix between completion rate and response rate at the condition level and the study level.

The parameter estimates (using RML with PQL estimation and second order Taylor linearization; see Chapter 6) are in Table 10.1. The first column shows the parameter estimates for the empty model. It produces two intercept estimates, indicated by 'comprate' and 'resprate', one for the completion rate and one for the response rate.

*Table 10.1* Results survey response data

| Model | M0: intercepts for comp. and resp. rate | M1: M0 + condition indicators |
|---|---|---|
| **Fixed part** | Coefficient (s.e.) | Coefficient (s.e.) |
| comprate | 0.84 (.13) | 1.15 (.16) |
| resprate | 1.28 (.15) | 1.40 (.16) |
| tel_comp | | –0.34 (.15) |
| tel_resp | | –0.10 (.11) |
| mail_comp | | –0.69 (.16) |
| mail_resp | | –0.40 (.13) |
| **Random part** | | |
| $\sigma^2_{comp/cond}$ | 0.41 (.09) | 0.31 (.07) |
| $\sigma^2_{resp/cond}$ | 0.20 (.05) | 0.18 (.04) |
| $r_{cr/cond}$ | 0.96 | 0.97 |
| $\sigma^2_{comp/cstudy}$ | 0.53 (.17) | 0.61 (.17) |
| $\sigma^2_{resp/study}$ | 0.89 (.22) | 0.83 (.21) |
| $r_{cr/study}$ | 0.99 | 0.95 |

Note that these estimates are not the same as the estimates we would get from two separate univariate analyses. If there is a tendency, for instance, to report only the response rate when the survey response is disappointing, because that looks better in the report, the omitted values for the completion rate are not missing completely at random. The univariate analysis of response rate has no way to correct for this bias; it assumes that any absent values are missing completely at random (MCAR). The multivariate model contains the covariance between the response rate and the completion rate. Hence, it can correct for the bias in reporting the response rate; it assumes that any absent values are missing at random (MAR), which is a weaker assumption. Because of this implicit correction, the intercepts and other regression coefficients in the multivariate model can be different from those estimated separately in univariate analyses. This is similar to the situation in multilevel longitudinal modeling (see Chapter 5), where panel dropout in the multilevel model is assumed to be missing at random (MAR). For an accessible discussion of the differences between MAR and MCAR see Allison (2002) and McKnight, McKnight, Sidani, and Figueredo (2007). As in multilevel longitudinal modeling, the fact that the multivariate multilevel model assumes that any absent outcome variables are MAR rather than MCAR is an important advantage when we have incomplete data. The usual practice in MANOVA to analyze only complete cases using listwise deletion assumes MCAR, which is a much stronger assumption than MAR.

The second column in Table 10.1 shows the parameter estimates for the model where the dummy variables that indicate the data collection conditions are added separately for the completion rate and the response rate. The face-to-face condition is the reference category, and two dummy variables are added that indicate the telephone and mail condition. We do not assume that the effect of the conditions is the same for both completion and response rate. Therefore, the two condition dummies are entered as interactions with the dummies that indicate the completion and response rate. Thus, the model equation is:

$$P_{hij} = \text{logistic}\begin{pmatrix} \gamma_{01}comp_{ij} + \gamma_{02}resp_{ij} \\ + \gamma_{03}tel_{ij}comp_{ij} + \gamma_{04}tel_{ij}resp_{ij} + \gamma_{05}mail_{ij}comp_{ij} + \gamma_{06}mail_{ij}resp_{ij} \\ + u_{1ij}comp_{ij} + u_{2ij}resp_{ij} + u_{1j}comp_{ij} + u_{2j}resp_{ij} \end{pmatrix}. \quad (10.15)$$

The two 'intercepts' for the completion rate and the response rate, and the regression slopes for the effect of the telephone and mail condition on the completion rate and the response rate seem to be quite different in Table 10.1. We can formally test the null-hypothesis that they are equal by testing the appropriate contrast. Testing the intercepts of *comp* and *resp* for equality, using the procedures described in Chapter 3, produces a chi-square of 6.82, which with one degree of freedom has a *p*-value of .01. Since the face-to-face condition is the reference category for the dummy variables, this gives a test of the equality of completion rate and response rate in the face-to-face condition. The same test produces for the telephone condition dummy variables a chi-square of 6.81, with one degree of freedom and a $p$-value of .01. For the mail condition, we get a chi-square of 8.94, with one degree of freedom and a $p$-value of .00. Clearly, the different data collection conditions affect the completion rate and the response rate in a different way.

The variance components are indicated in Table 10.1 by $\sigma^2_{comp/cond}$ for the intercept variance for the completion rate on the condition level, and $\sigma^2_{comp/study}$ for the intercept variance for the completion rate on the study level. Likewise, $\sigma^2_{resp/cond}$ indicates the intercept variance for the response rate on the condition level, and $\sigma^2_{resp/study}$ the intercept variance for the response rate on the study level. Note that Table 10.1 does not give a value for the deviance. The estimation is based on the quasi-likelihood approach described in Chapter 6, and therefore the deviance is approximate. For that reason, it is not included in the table.

If we add the explanatory variables publication year and saliency of survey topic, contrast tests show that these have similar effects on both the completion rate and the response rate. As a result, we can either add them to the regression equation as interactions with the completion and response rate dummies, constraining the equivalent regression slopes to be equal (see equations 10.9–10.11), or as a direct effect of the explanatory variables year and saliency (see equation 10.12).

Table 10.2 presents the parameter estimates for both model specifications. Both specifications produce the same value for the parameter estimates and the corresponding standard errors for the explanatory variables 'year' and 'saliency'. The model that includes the explanatory variables directly is given by:

$$P_{hij} = \text{logistic}\begin{pmatrix} \gamma_{01}comp_{ij} + \gamma_{02}resp_{ij} \\ + \gamma_{03}tel_{ij}comp_{ij} + \gamma_{04}tel_{ij}resp_{ij} + \gamma_{05}mail_{ij}comp_{ij} + \gamma_{06}mail_{ij}resp_{ij} \\ + \gamma_{07}year_j + \gamma_{08}saliency_j \\ + u_{1ij}comp_{ij} + u_{2ij}resp_{ij} + u_{1j}comp_{ij} + u_{2j}resp_{ij} \end{pmatrix}. \quad (10.16)$$

The model that includes these explanatory variables as interactions including two equality constraints, indicated by the superscripts $a$ and $b$, is given by:

$$P_{hij} = \text{logistic}\begin{pmatrix} \gamma_{01}comp_{ij} + \gamma_{02}resp_{ij} \\ + \gamma_{03}tel_{ij}comp_{ij} + \gamma_{04}tel_{ij}resp_{ij} + \gamma_{05}mail_{ij}comp_{ij} + \gamma_{06}mail_{ij}resp_{ij} \\ + \gamma^a_{07}year_jcomp_{ij} + \gamma^a_{08}year_jresp_{ij} \\ + \gamma^b_{09}saliency_jcomp_{ij} + \gamma^b_{10}saliency_jresp_{ij} \\ + u_{1ij}comp_{ij} + u_{2ij}resp_{ij} + u_{1j}comp_{ij} + u_{2j}resp_{ij} \end{pmatrix}. \quad (10.17)$$

Table 10.2 shows empirically what is derived in equations 10.9–10.12, namely that the two representations are equivalent. Since adding year and saliency directly is simpler, this is the preferred method.

*Table 10.2* Results survey response data, model comparison

| Model | Year and saliency as interaction terms | Year and saliency directly |
|---|---|---|
| **Fixed part** | Coefficient (s.e.) | Coefficient (s.e.) |
| comprate | 0.83 (.43) | 0.83 (.43) |
| resprate | 1.06 (.43) | 1.06 (.43) |
| tel_comp | –0.32 (.15) | –0.32 (.15) |
| tel_resp | –0.41 (.11) | –0.41 (.11) |
| mail_comp | –0.71 (.16) | –0.71 (.16) |
| mail_resp | –0.40 (.13) | –0.40 (.13) |
| year_comp[a] | –0.01 (.01) | n/a |
| year_resp[a] | –0.01 (.01) | n/a |
| sali_comp[b] | 0.69 (.17) | n/a |
| sali_resp[b] | 0.69 (.17) | n/a |
| year | n/a | –0.01 (.01) |
| saliency | n/a | 0.69 (.17) |
| **Random part** | | |
| $\sigma^2_{comp/cond}$ | 0.31 (.07) | 0.31 (.07) |
| $\sigma^2_{resp/cond}$ | 0.18 (.04) | 0.18 (.04) |
| $r_{cr/cond}$ | 0.97 | 0.97 |
| $\sigma^2_{comp/cstudy}$ | 0.45 (.14) | 0.45 (.14) |
| $\sigma^2_{resp/study}$ | 0.52 (.14) | 0.52 (.14) |
| $r_{cr/study}$ | 0.91 | 0.91 |

[a,b] Slopes constrained to be equal.

If we have a number of outcomes, all related to a single theoretical construct or syndrome, directly adding an explanatory variable to the model results in a higher power than adding them as a set of interactions with all outcome variables. The reason is that in the former case we use a one-degree of freedom test, and in the latter a *p*-degrees of freedom overall test. Adding an explanatory variable directly assumes that all interactions result in the same regression weight, which can subsequently be constrained to be equal. This assumption of a common effect size is strong, and it is not realistic if outcome variables are measured on different scales. Sammel et al. (1999) discuss the possibility of smoothing the regression coefficients. They suggest scaling the outcome variables prior to the analysis in such a way that they are measured on the same scale. With continuous variables, a transformation to standardized scores is appropriate. Raudenbush, Rowan, and Kang (1991) employ a transformation to correct for differences in measurement reliability. To arrive at comparable effect sizes this means that the outcomes are divided by the square root of the reliability coefficient, and then standardized.

## 10.3 EXAMPLE OF MULTIVARIATE MULTILEVEL ANALYSIS: MEASURING GROUP CHARACTERISTICS

Sometimes the interest may be in measuring characteristics of the context, that is, of the higher-level units, which can be individuals, groups, or organizations. For instance, we may be interested in school climate, and use a questionnaire that is answered by a sample of pupils from each of the schools. In this example we are not necessarily interested in the pupils, they are just used as informants to judge the school climate. Similar situations arise in health research, where patients may be used to express their satisfaction with their general practitioner, and community research, where samples from different neighborhoods evaluate various aspects of the neighborhood in which they live. In these cases, we may use individual characteristics to control for possible measurement bias, but the main interest is in measuring some aspect of the higher-level unit (see Paterson, 1998; Raudenbush & Sampson, 1999a; Sampson, Raudenbush, & Earls, 1997). This type of measurement was called 'ecometrics' by Raudenbush and Sampson (1999a).

Our example concerns data from an educational research study by Krüger (1994). In this study, male and female school managers were compared on a large number of characteristics. As part of the study, small samples of pupils from each school rated their school manager on six 7-point items that indicate a people-oriented approach toward leadership (the data are described in more detail in Appendix A). There are ratings from 854 pupils within 96 schools, 48 with a male and 48 with a female school manager, on these six items. If we calculate the reliability coefficient, Cronbach's alpha, for the six items, we get a reliability of .80, which is commonly considered sufficient to sum the items to a scale (Nunnally & Bernstein, 1994). However, this reliability estimate is difficult to interpret, because it is based on a mixture of school-level and individual pupil-level variance. Since all judgments within the same school are ratings of the same school manager, within-school variance does not give us information about the school manager. From the measurement point of view, we want to concentrate only on the between-schools variance.

One convenient way to model data such as these is to use a multivariate multilevel model, with separate levels for the items, the pupils, and the schools. Thus, we create six dummy variables to indicate the six items, and exclude the intercept from the model. Hence, at the lowest level we have:

$$Y_{hij} = \pi_{1ij} d_{1ij} + \pi_{2ij} d_{2ij} + \ldots + \pi_{6ij} d_{6ij}. \tag{10.18}$$

At the individual level we have:

$$\pi_{pij} = \beta_{pj} + u_{pij}. \tag{10.19}$$

At the group level (the third level in the multivariate model), we have:

$$\beta_{pj} = \gamma_p + u_{pj}. \tag{10.20}$$

By substitution, we obtain the single-equation version:

$$\begin{aligned} Y_{hij} = {} & \gamma_1\, d_{1ij} + \gamma_2 d_{2ij} + \ldots + \gamma_6 d_{pij} \\ & + u_{1ij} d_{1ij} + u_{2ij} d_{2ij} + \ldots + u_{6ij} d_{pij}. \\ & + u_{1j} d_{1ij} + u_{2j} d_{2ij} + \ldots + u_{6j} d_{pij} \end{aligned} \tag{10.21}$$

Using sum notation, we have:

$$Y_{hij} = \sum_{h=1}^{6} \gamma_h d_{hij} + \sum_{h=1}^{6} u_{hij} d_{hij} + \sum_{h=1}^{6} u_{hj} d_{hij}. \tag{10.22}$$

The model described by equations 10.21 and 10.22 provides us with estimates of the six item means, and of their variances and covariances at the pupil and school level. Since in this application we are mostly interested in the variances and covariances, restricted maximum likelihood (RML) estimation is preferred to full maximum likelihood (FML) estimation.

*Table 10.3* Covariances and correlations at the pupil level

| Item | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Item 1 | 1.19 | *.57* | *.44* | *.18* | *.25* | *.44* |
| Item 2 | 0.67 | 1.13 | *.52* | *.18* | *.26* | *.38* |
| Item 3 | 0.49 | 0.57 | 1.07 | *.19* | *.23* | *.43* |
| Item 4 | 0.17 | 0.17 | 0.17 | 0.74 | *.60* | *.30* |
| Item 5 | 0.22 | 0.23 | 0.20 | 0.42 | 0.66 | *.38* |
| Item 6 | 0.48 | 0.41 | 0.45 | 0.26 | 0.31 | 1.00 |

Note: The italic entries in the upper diagonal are the correlations.

*Table 10.4* Covariances and correlations at the school level

| Item | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Item 1 | 0.24 | *.91* | *.87* | *.57* | *.93* | *.96* |
| Item 2 | 0.30 | 0.45 | *.98* | *.14* | *.58* | *.88* |
| Item 3 | 0.24 | 0.36 | 0.31 | *.07* | *.53* | *.87* |
| Item 4 | 0.12 | 0.04 | 0.02 | 0.19 | *.89* | *.57* |
| Item 5 | 0.15 | 0.13 | 0.10 | 0.13 | 0.11 | *.90* |
| Item 6 | 0.16 | 0.20 | 0.17 | 0.09 | 0.10 | 0.12 |

Note: The italic entries in the upper diagonal are the correlations.

Table 10.3 presents the RML estimates of the covariances and the corresponding correlations at the pupil level, and Table 10.4 presents the same at the school level. The tables show that most of the variance of the six items is pupil-level variance, that is, variance between pupils within schools. Since within the same school all pupils are evaluating the same school manager, this variance must be regarded as systematic measurement bias. Apparently, the pupils differ systematically in the way they use the six items. The pattern of covariation in Table 10.3 shows how they differ. We can add pupil-level variables to the model, to investigate whether we can model this covariation. However, what we model in that case is individual idiosyncrasies in the way the measurement instrument is used by the pupils. From the perspective of measurement, we are mostly interested in Table 10.4, because this shows how the items perform at the school level. Although the variances at the school level are lower, the correlations are generally much higher. The mean correlation at the pupil level is .36, and at the school level .71. This is reassuring, because it means that at the school level the consistency of the measurement instrument is higher than at the individual level.

We can use the covariances or correlations in Table 10.4 to carry out an item analysis at the student or the school level, and we can use standard formulas from classical measurement theory to calculate the internal consistency reliability coefficient alpha. For instance, a convenient way to estimate the internal consistency given the results in Table 10.3 or 10.4 is to use the mean correlation (e.g., Nunnally & Bernstein, 1994). We can estimate the internal consistency of the scale from the mean correlation, using the Spearman-Brown formula for test length. With $p$ items, the reliability of the $p$-item scale is given by:

$$\alpha = p\bar{r}/(1 + (p - 1)\bar{r}) \tag{10.23}$$

where $\bar{r}$ is the mean correlation of the items, and $p$ is the scale length. The mean correlation at the school level is .71, and using the Spearman-Brown formula we can estimate the school-level coefficient alpha internal consistency as .94. However, this is not a very accurate estimate, since it ignores the differences in the variance of the items, but it produces a rough approximation. For a more accurate estimate, we could use the covariances in Table 10.3 or Table 10.4 as input in a software program for reliability or factor analysis, for a more formal analysis of the relationships between the items. If we do this, coefficient alpha is estimated as .92, and the item analysis further informs us that we should consider removing item 4 from the scale, because of its low correlations with the other items.

There is one important consideration. The matrix of covariances or correlations at the school level is a maximum likelihood estimator of the population matrix. It can be analyzed directly using models and software that can handle direct input of such matrices. Examples of using these matrices for multilevel structural equation analysis are given in Chapter 14. Using them for measurement at the school level is more questionable, because that assumes that we can actually observe the school-level residual errors that give rise to the covariances and correlations in Table 10.4. In actual fact, we cannot observe these residuals directly, let alone calculate their sum or mean. What we can observe is the school-level means of the evaluations given by the pupils in a specific school. Unfortunately, these school-level observed means also reflect the pupil-level variation; part of the variation at the school level is because of differences between pupils within schools. This issue will be taken up in detail in the chapters on multilevel structural equation modeling. In the context of multilevel measurement, it means that the observed school-level aggregated means contain error variation that is not visible in Table 10.4, so if their reliability is estimated using the covariances or correlations in Table 10.4 it will be overestimated.

Raudenbush et al. (1991) present an extensive discussion of the issues involved in multilevel measurement using observed group-level means. They provide (pp. 309–312) equations to calculate both the pupil-level and school-level internal consistency directly, using the intercept variances at the three available levels estimated in an intercept-only model. This model can be written as:

$$Y_{hij} = \gamma_{000} + u_{0hij} + u_{0ij} + u_{0j} \tag{10.24}$$

The model in equation 10.24 is the intercept-only model with three levels: the item, pupil, and school level. For our example, the variances are shown in Table 10.5, using an obvious notation for the subscripts of the variance components.

In Table 10.5, $\sigma^2_{item}$ can be interpreted as an estimate of the variation as a result of item inconsistency, $\sigma^2_{pupil}$ as an estimate of the variation of the scale score (mean item score) between different pupils within the same school, and $\sigma^2_{school}$ as an estimate of the

*Table 10.5* Intercept and variances for school manager data

| **Fixed part** | Coefficient | Standard error |
|---|---|---|
| Intercept | 2.57 | .05 |
| **Random part** | | |
| $\sigma^2_{school}$ | 0.179 | .03 |
| $\sigma^2_{pupil}$ | 0.341 | .03 |
| $\sigma^2_{item}$ | 0.845 | .02 |

variation of the scale score between different schools. These variances can be used to produce the internal consistency reliability at the pupil and school level. If we have *p* items, the error variance in the scale score (computed as the mean of the items) is given by $\sigma^2_e = \sigma^2_{item}/p = 0.845/6 = 0.141$.

The item level exists only to produce an estimate of the variance because of item inconsistency. We are in fact using a scale score that is computed as the mean of the items. The intraclass correlation of the scale score for the schools is given by $\rho_I = \sigma^2_{school}/(\sigma^2_{school} + \sigma^2_{pupil})$, which for our example is 0.179/(0.179 + 0.341) = .344. Therefore, for the scale score, about 34% of the variance is between schools.

The pupil-level internal consistency is estimated by $\alpha_{pupil} = \sigma^2_{pupil}/(\sigma^2_{pupil} + \sigma^2_{item}/p)$. For our example data this gives $\alpha_{pupil}$ = 0.341/(0.341 + 0.845/6) = .71. This reflects consistency in the variability of the ratings of the same school manager by pupils in the same schools. This internal consistency coefficient indicates that the pupil-level variability is not random error, but is systematic. It could be just systematic error, for instance response bias such as a halo effect in the judgments made by the pupils, or it could be based on different experiences of pupils with the same manager. This could be explored further by adding pupil characteristics to the model.

The school-level internal consistency is (Raudenbush et al., 1991, p. 312):

$$\alpha_{school} = \sigma^2_{school} / [\sigma^2_{school} + \sigma^2_{pupil} / n_j + \sigma^2_{item} / (p \cdot n_j)] \tag{10.25}$$

In equation 10.25, *p* is the number of items in the scale, and $n_j$ is the number of pupils in school *j*. Since the number of pupils varies across schools, the school-level reliability also varies. In schools with a larger sample of pupils the management style is measured more accurately than in schools with a small sample of pupils. To obtain an estimate of the average reliability across all schools, Raudenbush et al. (1991, p. 312) suggest using the mean of the schools' internal consistencies as a measure of the internal consistency reliability. A simpler approach is to use equation 10.25 with the mean number of pupils across all schools for $n_j$. In our example we have on average 8.9 pupils in each school, and if that number is plugged into equation 10.25 the overall school-level internal consistency is estimated as: $\alpha_{school} = 0.179/[0.179 + 0.341/8.9 + 0.845/(8.9 \times 6)] = .77$. The value of .77 for the school-level internal consistency coefficient indicates that the school managers' leadership style is measured with sufficient consistency.[2] The number of pupils per class varies between 4 and 10. If we plug these values into the equation, we find a reliability of .60 with 4 pupils, and .91 with 10 pupils. It appears that we need at least 4 pupils in each school to achieve sufficient measurement precision, as indicated by the school-level coefficient of internal consistency.

The school-level internal consistency depends on four factors: the number of items in the scale $k$, the mean correlation between the items on the school level $\bar{r}$, the number of pupils sampled in the schools $n_j$, and the intraclass correlation at the school level $\rho_I$. The school-level reliability as a function of these quantities is:

$$\alpha_{school} = \frac{kn_j\rho_I\bar{r}}{kn_j\rho_I\bar{r} + [(k-1)\bar{r} + 1](1-\rho_I)}. \tag{10.26}$$

The mean item intercorrelation at the school level can conveniently be estimated using the variances in the intercept-only model by $\bar{r} = \sigma^2_{pupil}/(\sigma^2_{pupil} + \sigma^2_{item})$.

Equation 10.26 shows that the internal consistency reliability can be improved by including more items in the scale, but also by taking a larger sample of pupils in each school. Raudenbush et al. (1991) demonstrate that increasing the number of pupils in the schools increases the school-level reliability faster than increasing the number of items in the scale. Even with a low inter-item correlation and a low intraclass correlation, increasing the number of pupils to infinity (admittedly hard to do) will in the end produce a reliability equal to one, whereas increasing the number of items to infinity will in general not.

In an analysis presented by Raudenbush et al. (1991), the measurement model is extended by combining items from several different scales in one analysis. The constant in the multilevel model is then replaced by a set of dummy variables that indicate to which scale each item belongs. This is similar to a confirmative factor analysis, with the restriction that the loadings of all items that belong to the same scale are equal, and that there is one common error variance. These are strong restrictions, which are often expressed as the assumption that the items are parallel (Lord & Novick, 1968). The usual assumptions for the internal consistency index are considerably weaker. For a multivariate analysis of complex relationships on a number of distinct levels,

[2] The difference from the estimate of .92 obtained using classical psychometric methods earlier reflects the fact that here we take the variation at the item and pupil level into account when we estimate the school-level reliability. The method presented here is more accurate.

multilevel structural equation modeling is both more powerful and less restrictive. These models are discussed in detail in Chapter 14. Note that the covariance matrices at the pupil and school level in Tables 10.3 and 10.4 could be used as input for such modeling.

If we want to predict the evaluation scores of the school manager using school-level variables, for instance the experience or gender of the school manager, or type of school, we can simply include these variables as explanatory variables in the multilevel model. Sometimes it is useful to have actual evaluation scores, for instance if we want to use these as explanatory variables in a different type of model. We can estimate the school managers' evaluation scores using the school-level residuals from the intercept-only model. Since these are centered on the school mean, the school mean must be added again to these residuals, to produce so-called posterior means for the evaluation scores. Since the posterior means are based on the empirical Bayes residuals, they are not simply the observed mean evaluation scores in the different schools, but they are shrunken toward the overall mean. The amount each score is shrunken toward the overall mean depends on the reliability of that score, which depends among other things on the number of pupils used in that particular school. The result is that we are using an estimate of the school-level true score of each school manager (see Lord & Novick, 1968; Nunnally & Bernstein, 1994). We can add pupil-level explanatory variables to the model, which would lead to evaluation scores that are conditional on the pupil-level variables. This can be used to correct the evaluation scores for inequalities in the composition of the pupil population across schools, which is important if the schools attract different types of students.

A nice feature of using multilevel modeling for measurement scales is that it accommodates incomplete data in a straightforward manner. If some of the item scores for some of the pupils are missing, this is compensated in the model. The model results and estimated posterior means are the correct ones, under the assumption that the data are missing at random (MAR). This is a weaker assumption than the missing completely at random (MCAR) assumption that is required with simpler methods, such as using only complete cases or replacing missing items by the mean of the observed items.

The measurement procedures just outlined are based on classical test theory, which means that they assume continuous multivariate normal outcomes. Most test items are categorical. If the items are dichotomous, we can use the logistic multilevel modeling procedures described in Chapter 6. Kamata (2001) shows that the two-level multilevel logistic model is equivalent to the Rasch model (Andrich, 1988), and discusses extensions to three-level models. If we have items with more than two categories, an ordinal multilevel can be used. Adams, Wilson, and Wu (1997) and Rijmen, Tuerlinckx, de Boeck, and Kuppens (2003) show how these models are related to item-response theory (IRT) models in general. In the interest of accurate measurement, maximum likelihood with numerical integration (see Chapter 6) or Bayesian estimation procedures (see Chapter 13) are preferable, especially with dichotomous items.

For general multivariate modeling of hierarchical data, multilevel structural equation modeling is more flexible than the multivariate multilevel regression model. These models are discussed in Chapters 14 and 15.

# 11

# The Multilevel Approach to Meta-Analysis

Meta-analysis is a systematic approach towards summarizing the findings of a collection of independently conducted studies on a specific research problem. In meta-analysis, statistical analyses are carried out on the published results of empirical studies on a specific research question. This chapter shows that multilevel regression models are attractive for analyzing meta-analytic data.

## 11.1 META-ANALYSIS AND MULTILEVEL MODELING

Meta-analysis is a systematic approach towards the synthesis of a large number of results from empirical studies (see Glass, 1976; Lipsey & Wilson, 2001). The goal is to summarize the findings of a collection of independently conducted studies on a specific research problem. For instance, the research question might be: 'What is the effect of social skills training on socially anxious children?' In a meta-analysis, one would collect reports of experiments concerning this question, explicitly code the reported outcomes, and integrate the outcomes statistically into a combined 'super outcome'. Often the focus is not so much on integrating or summarizing the outcomes, but on more detailed questions such as: 'What is the effect of different durations for the training sessions?' or 'Are there differences between different training methods?' These questions address the generalizability of the research findings. In these cases, the meta-analyst not only codes the study outcomes, but also codes study characteristics. These study characteristics are potential explanatory variables to explain differences in the study outcomes. Meta-analysis is not just the collection of statistical methods used to achieve integration; it is the application of systematic scientific strategies to the literature review (Cornell & Mulrow, 1999; Light & Pillemer, 1984). For a brief introduction to general meta-analysis I refer to Cornell and Mulrow (1999), and Lipsey and Wilson (2001). A thorough and complete treatment of methodological and statistical issues in meta-analysis, including a chapter on using multilevel regression methods (Raudenbush, 2009), can be found in Cooper, Hedges, and Valentine (2009) and in Sutton, Abrams, Jones, Sheldon, and Song (2000).

The core of meta-analysis is that statistical analyses are carried out on the published results of a collection of empirical studies on a specific research question. One approach is to combine the *p*-values of all the collected studies into one combined *p*-value. This is a simple matter, but does not provide much information. A very general model for meta-analysis is the random-effects model (Hedges & Olkin, 1985, p. 198). In this model, the focus is not on establishing the statistical significance of a combined outcome, but on analyzing the variation of the effect sizes found in the different studies. The random-effects model for meta-analysis assumes that study outcomes vary across studies, not only because of random sampling effects, but also because there are real differences between the studies. For instance, study outcomes may vary because the different studies employ different sampling methods, use different experimental manipulations, or measure the outcome with different instruments. The random-effects model is used to decompose the variance of the study outcomes into two components: one component that is the result of sampling variation, and a second component that reflects real differences between the studies. Hedges and Olkin (1985) and Lipsey and Wilson (2001) describe procedures that can be used to decompose the total variance of the study outcomes into random sampling variance and systematic between-studies variance, and procedures to test the significance of the between-studies variance. If the between-studies variance is large and significant, the study outcomes are regarded as *heterogeneous*. This means that the studies do not all provide the same outcome. One procedure to investigate the differences between the studies is to form clusters of studies which differ in their outcomes between the clusters, but are homogeneous within the clusters. The next goal is to identify study characteristics that explain differences between the clusters. Variables that affect the study outcomes are in fact moderator variables: that is, variables that interact with the independent variable to produce variation in the study outcomes.

Meta-analysis can be viewed as a special case of multilevel analysis. We have a hierarchical data set, with subjects within studies at the first level, and studies at the second level. If the raw data of all the studies were available, we could carry out a standard multilevel analysis, predicting the outcome variable using the available individual- and study-level explanatory variables. In our example on the effect of social skills training of children, we would have one outcome variable, for instance the result on a test measuring social skill, and one explanatory variable, which is a dummy variable, that indicates whether the subject is in the experimental or the control group. At the individual level, we have a linear regression model that relates the outcome to the experimental/control-group variable. The general multilevel regression model assumes that each study has its own regression model. If we have access to all the original data, standard multilevel analysis can be used to estimate the mean and variance of the regression coefficients across the studies. If the variance of the regression slopes of the experimental/control-group variable is large and significant, we have heterogeneous results. In that case, we can use the available study characteristics as explanatory variables at the second (study) level to predict the differences of the regression coefficients.

These analyses can be carried out using standard multilevel regression methods and standard multilevel software. However, in meta-analysis we usually do *not* have access to the original raw data. Instead, we have the published results in the form of *p*-values, means, standard deviations, or correlation coefficients. Classical meta-analysis has developed a large variety of methods to integrate these statistics into one overall outcome, and to test whether these outcomes should be regarded as homogeneous or heterogeneous. Hedges and Olkin (1985) discuss the statistical models on which these methods are based. They describe a weighted regression model that can be used to model the effect of study characteristics on the outcomes, and Lipsey and Wilson (2001) show how conventional software for weighted regression analysis can be used to analyze meta-analytic data. Hunter and Schmidt (2004) discuss different approaches, but also use random-effects models that resemble multilevel analysis.

Even without access to the raw data, it is often possible to carry out a multilevel meta-analysis on the summary statistics that are the available data for the meta-analysis. Raudenbush and Bryk (Raudenbush & Bryk, 1985, 2002) view the random-effects model for meta-analysis as a special case of the multilevel regression model. The analysis is performed on sufficient statistics instead of raw data, and as a result some specific restrictions must be imposed on the model. In multilevel meta-analysis, it is simple to include study characteristics as explanatory variables in the model. If we have hypotheses about study characteristics that influence the outcomes, we can code these and include them on a priori grounds in the analysis. Alternatively, after we have concluded that the study outcomes are heterogeneous, we can explore the available study variables in an attempt to explain the heterogeneity.

The major advantage of using multilevel analysis instead of classical meta-analysis methods is flexibility (Hox & de Leeuw, 2003). Using a multilevel framework, it is easy to add further levels to the model, for example to accommodate multiple outcome variables. Estimation can be done using maximum likelihood methods, and a range of estimation and testing methods are available (see Chapters 3 and 13). However, not all multilevel analysis software can be used for multilevel meta-analysis; the main requirement is that it is possible to impose constraints on the random part of the model.

## 11.2 THE VARIANCE-KNOWN MODEL

In a typical meta-analysis, the collection of studies found in the literature employ different instruments and use different statistical tests. To make the outcomes comparable, the study results must be transformed into a standardized measure of the effect size, such as a correlation coefficient or the standardized difference between two means. For instance, if we perform a meta-analysis on studies that compare an experimental group to a control group, an appropriate measure for the effect size is the standardized difference between two means *g*, which is given by $g = (\bar{Y}_E - \bar{Y}_C)/s$. The standard deviation *s* is either the standard deviation in the control group, or the pooled standard deviation for both the experimental and the control group. Since the standardized difference *g* has a small upwards bias, it is often transformed to the unbiased effect size indicator $d = (1 - 3/(4N - 9))g$, where *N* is the total sample size for the study. This correction is most appropriate when *N* is less than 20; with larger sample sizes the bias correction is often negligible (Hedges & Olkin, 1985).

The general model for the study outcomes, ignoring possible study-level explanatory variables, is given by:

$$d_j = \delta_j + e_j \tag{11.1}$$

In equation 11.1, $d_j$ is the outcome of study *j* ($j = 1, \ldots, J$), $\delta_j$ is the corresponding population value, and $e_j$ is the sampling error for this specific study. It is assumed that the $e_j$ have a normal distribution with a known variance $\sigma_j^2$. If the sample sizes of the individual studies are not too small, for instance between 20 (Hedges & Olkin, 1985, p. 175) and 30 (Raudenbush & Bryk, 2002, p. 207), it is reasonable to assume that the sampling distribution of the outcomes is normal, and that the known variance can be estimated from the data with sufficient accuracy. The assumption of underlying normality is not unique for multilevel meta-analysis; most classical meta-analysis methods also assume normality (see Hedges & Olkin, 1985). The variance of the sampling distribution of the outcome measures is assumed to be known from statistical theory.

To obtain a good approximation to a normal sampling distribution, and to determine the known variance, a transformation of the original effect size statistic is often needed. For instance, since the sampling distribution of a standard deviation is only approximately normal, it should not be used with small samples. The transformation $s^* = \ln(s) + 1/(2df)$ of the standard deviation improves the normal approximation. The usual transformation for the correlation coefficient *r* is the familiar Fisher-*Z* transformation, and for the proportion it is the logit. Note that, if we need to perform a meta-analysis on logits, the procedures outlined in Chapter 6 are generally more accurate. Usually, after a confidence interval is constructed for the transformed variable, the end-points are translated back to the scale of the original estimator. Table 11.1 lists some common effect size measures, the usual transformation if one is needed, and the sampling variance (of the transformed outcome if applicable) (Bonett, 2002; Lipsey & Wilson, 2001; Raudenbush & Bryk, 2002; Rosenthal, 1994).

Equation 11.1 shows that the effect sizes $\delta_j$ are assumed to vary across the studies. The variance of the $\delta_j$ is explained by the regression model:

$$\delta_j = \gamma_0 + \gamma_1 Z_{1j} + \gamma_2 Z_{2j} + \ldots + \gamma_p Z_{pj} + u_j, \tag{11.2}$$

*Table 11.1* Some effect measures, their transformation and sampling variance

| Measure | Estimator | Transformation | Sampling variance |
|---|---|---|---|
| Mean | $\bar{x}$ | – | $s^2/n$ |
| Diff. 2 means | $g = (\bar{y}_E - \bar{y}_C)/s$ | $d = (1 - 3/(4N - 9))g$ | $(n_E + n_C)/(n_E n_C) + d^2/(2(n_E + n_C))$ |
| Stand. dev. | $s$ | $s^* = \ln(s) + 1/(2df)$ | $1/(2df)$ |
| Correlation | $r$ | $z = 0.5\ln((1 + r)/(1 - r))$ | $1/(n - 3)$ |
| Proportion | $p$ | $\text{logit} = \ln(p/(1 - p))$ | $1/(np(1 - p))$ |
| Diff. 2 prop. | $d \approx Z_{p1} - Z_{p2}$ | – | $\frac{2\pi p_1(1 - p_1)e^{Z_{p1}^2}}{n_1} + \frac{2\pi p_2(1 - p_2)e^{Z_{p2}^2}}{n_2}$ |
| Diff. 2 prop. | $\text{logit}(p_1) - \text{logit}(p_2)$ | – | $\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$ ($a$, $b$, $c$, $d$ are cell freq.) |
| Reliability = coefficient alpha | Cronbach's α | $Z = \ln(1 - \lvert a \rvert)$ | $\frac{k}{2(k - 1)(n - 2)}$ ($k$ = # of items) |

where $Z_1 \ldots Z_p$ are study characteristics, $\gamma_1, \ldots, \gamma_p$ are the regression coefficients, and $u_j$ is the residual error term, which is assumed to have a normal distribution with variance $\sigma_u^2$. By substituting 11.2 into 11.1 we obtain the complete model:

$$d_j = \gamma_0 + \gamma_1 Z_{1j} + \gamma_2 Z_{2j} + \ldots + \gamma_p Z_{pj} + u_j + e_j. \tag{11.3}$$

If there are no explanatory variables, the model reduces to:

$$d_j = \gamma_0 + u_j + e_j. \tag{11.4}$$

Model 11.4, which is the 'intercept-only' or 'empty' model, is equivalent to the random-effects model for meta-analysis described by Hedges and Olkin (1985).

In model 11.4, the intercept $\gamma_0$ is the estimate for the mean outcome across all studies. The variance of the outcomes across studies, $\sigma_u^2$, indicates how much these outcomes vary across studies. Thus, testing if the study outcomes are homogeneous is equivalent to testing the null-hypothesis that the variance of the residual errors $u_j$, indicated by $\sigma_u^2$, is equal to zero. If the test of $\sigma_u^2$ is significant, the study outcomes are considered heterogeneous. The proportion of systematic between-study variance can be estimated by the intraclass correlation $\rho = \sigma_u^2/(\sigma_u^2 + \sigma_e^2)$.

The general model 11.3 includes study characteristics $Z_{pj}$ to explain differences in the studies' outcomes. In model 11.3, $\sigma_u^2$ is the residual between-study variance after the explanatory variables are included in the model. A statistical test on $\sigma_u^2$ now tests whether the explanatory variables in the model explain all the variation in the studies' outcomes, or if there still is unexplained systematic variance left in the outcomes. The difference between the between-studies variance $\sigma_u^2$ in the empty model and in the model that includes the explanatory variables $Z_{pj}$ can be interpreted as the amount of variance explained by the explanatory variables, that is, by the study characteristics.

The multilevel meta-analysis model given by equation 11.3 is equal to the general weighted regression model for random effects described by Hedges and Olkin (1985, Chapter 9). When the study-level variance is not significant, it can be removed from the model:

$$d_j = \gamma_0 + \gamma_1 Z_{1j} + \gamma_2 Z_{2j} + \ldots + \gamma_p Z_{pj} + e_j. \tag{11.5}$$

Compared to model 11.3, model 11.5 lacks the study-level residual error term $u_j$. The result is called the fixed-effect model, which assumes that all studies are homogeneous and all estimate the same underlying population parameter $\delta$. Thus, the fixed-effect model described by Hedges and Olkin (1985, Chapter 8) is a special case of the random-effects weighted regression or the multilevel meta-analysis model. Omitting the study-level residual error term $u_j$ implies that there is no variation in the effect sizes across all studies, or that the explanatory variables in the model explain all the variance among the studies. Thus, if the residual between-study variance is zero, a fixed-effect model is appropriate (Hedges & Vevea, 1998). However, this assumption is not very realistic. For instance, Hunter and Schmidt (2004) argue that the between-studies heterogeneity is partly produced by some unavoidable artifacts encountered in meta-analysis. Examples of such artifacts are the (usually untestable) assumption of a normal distribution for the sampling errors $e_j$, the correctness of statistical assumptions made in the original analyses, differences in reliability of instruments used in different studies, coder unreliability, and so on. It is unlikely that the available study-level variables cover all these artifacts. Generally, the amount of detail in the input for the meta-analysis, the research reports, papers and articles, is not enough to code all these study characteristics for all of the studies. Therefore, heterogeneous results are to be expected (see Engels, Schmidt, Terrin, Olkin, & Lau, 2000). Since heterogeneous results are common, Hunter and Schmidt recommend as a rule of thumb that the study-level variance should be larger than 25% of all variance to merit closer inspection; study-level variance smaller than 25% is likely to be the result of methodological differences between the studies. However, simulations have shown that this '25% rule' is very inaccurate and therefore is not recommended (Schulze, 2008).

If fixed-effect models are used in the presence of significant between-study variance, the resulting confidence intervals are biased and much too small (Brockwell & Gordon, 2001; Villar, Mackey, Carroli, & Donner, 2001). If random-effects models are used, the standard errors are larger, and the estimate of the average effect may be different, depending on the relation between effect sizes and sample sizes in the primary studies (see Villar et al., 2001). Since there is, as a rule, unexplained variance in a meta-analysis, random-effects models are generally preferred over fixed-effect models. Lipsey and Wilson (2001) describe a weighted least squares regression procedure for estimating the model parameters, which can be applied using standard statistical software for weighted regression. Just like multilevel meta-analysis this is a powerful approach, because one can include explanatory variables in the model. However, in the weighted regression approach the investigators must supply an estimate of the between-study variance. This variance is estimated before the weighted regression analysis, and the estimated value is then plugged into the weighted regression analysis (Lipsey & Wilson, 2001). Multilevel analysis programs estimate this variance component, typically using iterative maximum likelihood estimation, which in general is more precise and efficient. In practice, both approaches usually produce very similar parameter estimates. The multilevel approach has the additional advantage that it offers more flexibility, for example, by using a three-level model for multivariate outcomes.

## 11.3 EXAMPLE AND COMPARISON WITH CLASSICAL META-ANALYSIS

In this section we analyze an example data set using classical meta-analysis methods as implemented in the meta-analysis macros written by David Wilson (Lipsey & Wilson, 2001, Appendix D). These macros are based on methods and procedures described by Hedges and Olkin (1985). The (simulated) data set consists of 20 studies that compare an experimental group and a control group.

Let us return to our example on the effect of social skills training on socially anxious children. We collect reports of experiments concerning this question. If we compare the means of an experimental and a control group, an appropriate outcome measure is the standardized difference between the experimental and the control group, originally proposed by Glass (1976) and defined by Hedges and Olkin as $g = (\bar{Y}_E - \bar{Y}_C)/s$, where $s$ is the pooled standard deviation of the two groups. Because $g$ is not an unbiased estimator of the population effect $\delta = (\mu_E - \mu_C)/\sigma$, Hedges and Olkin prefer a corrected effect measure $d$: $d = (1 - 3/(4N - 9))g$. The sampling variance of the effect estimator $d$ is $(n_E + n_C)/(n_E n_C) + d^2/(2(n_E + n_C))$ (Hedges & Olkin, 1985, p. 86).

Table 11.2 is a summary of the outcomes from a collection of 20 studies. The studies are presented in increasing order of their effect sizes (*g*, *d*). Table 11.2 presents both *g* and *d* for all 20 studies, with some study characteristics. The difference between *g* and *d* is very small in most cases, where study sample sizes are larger than about 20. Table 11.2 also presents the sampling variance of the effect sizes *d* (var(*d*)), the one-sided *p*-value of the *t*-test for the difference of the two means (*p*), the number of cases in the experimental ($n_{exp}$) and control group ($n_{con}$), and the reliability ($r_{ii}$) of the outcome measure used in the study. The example data set contains several study-level explanatory variables. A theoretically motivated explanatory variable is the duration in number of weeks of the experimental intervention. It is plausible to assume that longer interventions lead to a larger effect. In addition we have the reliability of the outcome measure ($r_{ii}$), and the size of the experimental and control group.

*Table 11.2* Example meta-analytic data from 20 studies

| Study | Weeks | $g$ | $d$ | var($d$) | $p$ | $n_{exp}$ | $n_{con}$ | $r_{ii}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 3 | −.268 | −.264 | .086 | .810 | 23 | 24 | .90 |
| 2 | 1 | −.235 | −.230 | .106 | .756 | 18 | 20 | .75 |
| 3 | 2 | .168 | .166 | .055 | .243 | 33 | 41 | .75 |
| 4 | 4 | .176 | .173 | .084 | .279 | 26 | 22 | .90 |
| 5 | 3 | .228 | .225 | .071 | .204 | 29 | 28 | .75 |
| 6 | 6 | .295 | .291 | .078 | .155 | 30 | 23 | .75 |
| 7 | 7 | .312 | .309 | .051 | .093 | 37 | 43 | .90 |
| 8 | 9 | .442 | .435 | .093 | .085 | 35 | 16 | .90 |
| 9 | 3 | .448 | .476 | .149 | .116 | 22 | 10 | .75 |
| 10 | 6 | .628 | .617 | .095 | .030 | 18 | 28 | .75 |
| 11 | 6 | .660 | .651 | .110 | .032 | 44 | 12 | .75 |
| 12 | 7 | .725 | .718 | .054 | .003 | 41 | 38 | .90 |
| 13 | 9 | .751 | .740 | .081 | .009 | 22 | 33 | .75 |
| 14 | 5 | .756 | .745 | .084 | .009 | 25 | 26 | .90 |
| 15 | 6 | .768 | .758 | .087 | .010 | 42 | 17 | .90 |
| 16 | 5 | .938 | .922 | .103 | .005 | 17 | 39 | .90 |
| 17 | 5 | .955 | .938 | .113 | .006 | 14 | 31 | .75 |
| 18 | 7 | .976 | .962 | .083 | .002 | 28 | 26 | .90 |
| 19 | 9 | 1.541 | 1.522 | .100 | .0001 | 50 | 16 | .90 |
| 20 | 9 | 1.877 | 1.844 | .141 | .00005 | 31 | 14 | .75 |

### 11.3.1 Classical meta-analysis

Classical meta-analysis includes a variety of approaches that complement each other. For instance, several different formulas are available for combining $p$-values. A classic procedure is the so-called Stouffer method (Rosenthal, 1991). In the Stouffer method, each individual (one-sided) $p$ is converted to the corresponding standard normal $Z$-score. The $Z$-scores are then combined using $Z = \left(\sum Z_j\right)/\sqrt{k}$, where $Z_j$ is the $Z$-value of study $j$, and $k$ is the number of studies. For our example data, the Stouffer method gives a combined $Z$ of 7.73, which is highly significant ($p < .0001$).

The combined $p$-value gives us evidence that an effect exists, but no information on the size of the experimental effect. The next step in classical meta-analysis is to combine the effect sizes of the studies into one overall effect size, and to establish the significance or a confidence interval for the combined effect. Considering the possibility that the effects may differ across the studies, the random-effects model is preferred to combine the studies.

In classical meta-analysis, the fixed-effect model is used first to combine the effect sizes. It is clear that larger studies include less sampling error, and therefore deserve a larger weight in combining the effect sizes. Hedges and Olkin (1985) prove that the optimal weight is not the sample size, but the precision, which is equal to the inverse of the sampling variance. The sample size and the inverse variance weight are obviously highly related. Hence, the fixed-effect model weights each study outcome with the inverse variance of the effect size: $w_j = 1/\text{var}(d_j)$. The combined effect size is simply the weighted mean of the effect sizes. The standard error of the combined effect size is calculated as the square root of the sum of the inverse variance weights:

$$SE_{\bar{d}} = \sqrt{\frac{1}{\sum w_j}}. \tag{11.6}$$

The test statistic to test for homogeneity of study outcomes is:

$$Q = \sum w_j \left(d_j - \bar{d}\right), \tag{11.7}$$

which has a chi-square distribution with $J - 1$ degrees of freedom. If the chi-square is significant, we reject the null-hypothesis of homogeneity and conclude that the studies are heterogeneous; there is significant study-level variation. In classical meta-analysis, the study-level variance is estimated by a method of moments estimator given by:

$$\sigma_u^2 = \frac{Q - (J - 1)}{\sum w_j - \left(\sum w_j^2 / \sum w_j\right)}. \tag{11.8}$$

The random-effects model follows the same procedures, but recalculates the weights by plugging in the estimate of the study-level variance:

$$w_j^* = \frac{1}{\text{var}(d_j) + \sigma_u^2}, \tag{11.9}$$

The random-effects model adds the between-study-level variance to the known variances when calculating the inverse variance weight. Subsequently, the same methods are used to estimate the combined effect size and its standard error.

A meta-analysis of the effect sizes in Table 11.2, using the random-effects model and the methods described earlier (using the macro MEANES), estimates the overall effect as $\delta = .580$, with a standard error of .106. Using this information, we can carry out a null-hypothesis test by computing $Z = d/SE(d) = 0.58/.106 = 5.47$ ($p < .0001$). The 95% confidence interval for the overall effect size is $0.37 < \delta < 0.79$. The usual significance test of the between-study variance used in meta-analysis is a chi-square test on the residuals, which for our example data leads to $\chi^2 = 49.59$ ($df = 19$, $p < .001$). This test is equivalent to the chi-square residuals test described by Raudenbush and Bryk (2002) and implemented in HLM. As the result is clearly significant, we have heterogeneous outcomes. This means that the overall effect 0.58 is not the estimate of a fixed population value, but an average of the distribution of effects in the population. The $Z$-value of 5.47 computed using the random-effects model is not the same as the $Z$-value of 7.73 computed using the Stouffer method. This difference is most likely because of a difference in power between these methods (Becker, 1994). Since the random-effects meta-analysis produces a standard error which can be used to establish a confidence interval, we will use the results from the meta-analysis.

The parameter variance $\sigma_u^2$ is estimated as 0.14, and the proportion of systematic variance is estimated as 0.65 (estimated as $\sigma_u^2$ divided by the weighted observed variance). This is much larger than the 0.25 that Hunter and Schmidt (2004) consider a lower limit for examining differences between studies. The conclusion is that the between-study variance is not only significant, but also large enough to merit further analysis using the study characteristics at our disposal. The usual follow-up in classical meta-analysis is to use weighted regression to analyze differences between study outcomes. When the random effects model is used, the same variance estimate described earlier is plugged into the calculation of the weight, and then weighted regression methods are used to estimate regression weights for the study-level variables. Instead of using the plug-in estimate, iterative maximum likelihood methods are also available, but they are less commonly used (see Lipsey & Wilson, 2001, p. 119). Using the method of moments estimator, the regression coefficient of the variable weeks is estimated as 0.14, with a standard error of .034 and an associated $p$-value of .0000. So the hypothesis that the intervention effect is larger with longer durations of the intervention is sustained. The result of the homogeneity test, conditional on the predictor weeks, is $Q$ = 18.34 ($df$ = 18, $p$ = .43). There is no evidence for study-level heterogeneity once the differences in duration are accounted for. The residual variance is $\sigma_u^2$ and is estimated as 0.04. The explained variance can be estimated as (0.14 – 0.04)/ 0.14 = 0.71. Given that the residual variance is not significant, one could decide to consider a fixed model where the variable weeks is included. However, the chi-square test for the between-study variance has a low power unless the number of studies is large (at least 50), so it is recommended to keep the between-studies variance in the model (see Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006).

### 11.3.2 Multilevel meta-analysis

A multilevel meta-analysis of the 20 studies using the empty 'intercept-only' model produces virtually the same results as the classical meta-analysis. Since in meta-analysis we are strongly interested in the size of the between-study variance component, restricted maximum likelihood (RML) estimation is the best approach.[1] Using RML, the intercept, which in the absence of other explanatory variables is the overall outcome, is estimated as $\gamma_0$ = 0.57, with a standard error of .11 ($Z$ = 5.12, $p$ < .001). The parameter variance $\sigma_u^2$ is estimated as 0.15 (s.e. = .111, $Z$ = 1.99, $p$ = .02). As the Wald test is inaccurate for testing variances (see Chapter 3), the variance is also tested with the deviance difference test. This produces a chi-square of 10.61 ($df$ = 1, halved $p$ < .001). The proportion of systematic variance is 0.71, which is much larger than 0.25, the lower limit for examining differences between studies (Hunter & Schmidt, 2004). The differences between these results and the results computed using the classical approach to meta-analysis are small, indicating that the classical approach is quite accurate when the goal of the meta-analysis is to synthesize the results of a set of studies.

When we include the duration of the experimental intervention as an explanatory variable in the regression model, we have:

$$d_j = \gamma_0 + \gamma_1 duration_{1j} + u_j + e_j \tag{11.10}$$

The results of the multilevel meta-analysis are summarized in Table 11.3, which presents the results for both the empty (null) model and the model that includes duration and the results obtained by the classical (random-effects) meta-analysis.

[1] If RML is used, it is not possible to test the effect of moderator variables using the deviance test. In practice, when the difference between FML and RML estimation is small, it may be advantageous to use FML rather than RML. If the differences are appreciable, RML is recommended.

The results are very close. It should be noted that the chi-square from the method of moments analysis is not straightforward: to obtain correct estimates and standard errors for the regression parameters in the second column we need to use a random-effects model with the plug-in variance estimate; to obtain the correct chi-square for the residual variance we must use the fixed-effect model. The multilevel analysis using the built-in meta-analysis option in HLM directly produces the chi-square residuals test. (Different choices and variations in software implementation are discussed in an appendix to this chapter.)

*Table 11.3* Results for random-effects method of moments and multilevel analyses

| Analysis | Method of moments | Method of moments | Multilevel RML | Multilevel RML |
|---|---|---|---|---|
| Delta/intercept | 0.58 (.11) | –0.22 (.21) | 0.58 (.11) | –0.23 (.21) |
| Duration | | 0.14 (.03) | | 0.14 (.04) |
| $\sigma_u^2$ | 0.14 | 0.04 | 0.15 (.08) | 0.04 (04) |
| $\chi^2$ deviance test & *p*-value | n/a | n/a | $\chi^2 = 10.6$<br>$p < .001$ | $\chi^2 = 1.04$<br>$p = .16$ |
| $\chi^2$ residuals test & *p*-value | $\chi^2 = 49.6$<br>$p < .001$ | $\chi^2 = 26.4$<br>$p = .09$ | $\chi^2 = 49.7$<br>$p < .001$ | $\chi^2 = 26.5$<br>$p = .09$ |

After including duration as an explanatory variable in the model, the residual between-study variance is much smaller, and no longer significant. The regression coefficient for the duration is 0.14 ($p$ < .001), which means that for each additional week the expected gain in study outcome is 0.14. The intercept in this model is –0.23, with a standard error of .21 ($p$ = .27). The intercept is not significant, which is logical, because it refers to the expected outcome of a hypothetical experiment with duration of zero weeks. If we center the duration variable by subtracting its overall mean, the intercept does not change from one model to the next, and reflects the expected outcome of the average study. The residual variance in the last model is 0.04, which is not significant. If we compare this with the parameter variance of 0.14 in the empty model, we conclude that 73% of the between-studies variance can be explained by including 'duration' as the explanatory variable in the model.

In the multilevel analyses reported in Table 11.3, RML estimation is used, and the residual between-studies variance is tested for significance twice, once using the deviance difference test, and once using the chi-square test proposed by Raudenbush and Bryk (2002). (The deviance test is not available in the method of moments.) As explained in more detail in Chapter 3, there are two reasons for choosing RML estimation and *not* using the Wald test on the variance. First, in standard applications of multilevel analysis, the variances are often viewed as nuisance parameters. It is important to include them in the model, but their specific value is not very important, because they are not interpreted. In meta-analysis, the question of whether all the studies report essentially the same outcome is an essential research question. The answer to this question depends on the size and on the decision about the significance of the between-studies variance. Therefore, it is very important to have a good estimate of the between-studies variance and its significance. For this reason, restricted maximum likelihood (RML) estimation is used instead of full maximum likelihood (FML). Generally, FML and RML estimation lead to very similar variance estimates, but if they do not, using RML provides better estimates (Browne, 1998). Second, the asymptotic Wald test on the variance computes the test statistic $Z$ by dividing the variance estimate by its standard error. This assumes a normal sampling distribution for the variance. This assumption is not justified, because variances are known to have a chi-square sampling distribution. Compared to other tests, the Wald test of the variance has a much lower power (Berkhof & Snijders, 2001), and in general the deviance difference test is preferred (Berkhof & Snijders, 2001; LaHuis & Ferguson, 2009). The difference between the deviance difference test and the residuals chi-square test is small, unless the group sample sizes are small. A practical reason for reporting the chi-square residuals test for the variance in a meta-analysis is that the residuals chi-square test proposed by Raudenbush and Bryk (2002) follows the same logic as the chi-square test on the residuals in classic meta-analysis, which facilitates comparison.

Since the study outcome depends in part on the duration of the experiment, reporting an overall outcome for the 20 studies does not convey all the relevant information. We could report the expected outcome for different duration, or calculate which duration is the minimum needed to obtain a significant outcome. This can be accomplished by centering the explanatory variable on different values. For instance, if we center the duration around 2 weeks, the intercept can be interpreted as the expected outcome at 2 weeks. Some multilevel analysis programs can produce predicted values with their expected error variance, which is also useful for describing the expected outcome for experiments with a different duration.

## 11.4 CORRECTING FOR ARTIFACTS

Hunter and Schmidt (2004) encourage correcting study outcomes for a variety of artifacts. It is common to correct the outcome for the attenuation that results from unreliability of the measure used. The correction simply divides the outcome measure by the square root of the reliability, for instance $d^* = d / \sqrt{r_{ii}}$, after which the analysis is carried out as usual. This is the same correction as the classical correction for attenuation of the correlation coefficient in psychometric theory (see Nunnally & Bernstein, 1994). Hunter and Schmidt (2004) describe many more corrections, but all of them share major methodological and statistical problems. One problem is that the majority of corrections always result in larger effect sizes. For instance, if the studies use instruments with a low reliability, the corrected effect size is much larger than the original effect size. If the reported reliability is incorrect, so will be the correction. Because the large effects have in fact not been observed, routinely carrying out such corrections is controversial. For that reason, Schwarzer (1989, p. 56) advises always to report the original values in addition to the corrected results. A second problem with all these corrections is that they influence the standard error of the outcome measure. Lipsey and Wilson (2001) present proper standard errors for some corrections. However, if the values used to correct outcomes are themselves subject to sampling error, the sampling variance of the outcome measure becomes still larger. Especially if many corrections are performed, their cumulative effect on the bias and accuracy of the outcome measures is totally unclear.

A different approach to correcting artifacts is to include them as covariates in the multilevel regression analysis. For reliability of the outcome measure, this is still not optimal, because the proper correction is a nonlinear multiplicative model (see Nunnally & Bernstein, 1994), and regression analysis is linear and additive. However, if the reliabilities are not extremely low (Nunnaly and Bernstein suggest as a rule of thumb that the reliability of a 'good' measure should be larger than .70), a linear additive model is a reasonable approximation, and we can always include quadratic or cubic trends in the analysis if that is needed. Figure 11.1 shows the effect of the correction for attenuation for $d = .5$ (medium effect size) and a reliability ranging between zero and one. It is clear that for reliabilities larger than .5 the relationship is almost linear.

The advantage of adding reliability as a predictor variable is that the effect of unreliability on the study outcomes is estimated based on the available data and not by a priori corrections. Another advantage is that we can test statistically if the correction has indeed a significant effect. Lastly, an interesting characteristic of multilevel modeling in meta-analysis is that it is possible to add an explanatory variable only to the random part, excluding it from the fixed part. Hence, if we suspect that a certain covariate, for instance poor experimental design, has an effect on the variability of the outcomes, we have the option to include it only in the random part of the model, where it affects the between-studies variance, but not the average outcome.

A variation on correcting for artifacts is controlling for the effect of study size. An important problem in meta-analysis is the so-called *file drawer problem*. The data for a meta-analysis are the results from previously published studies. Studies that find significant results may have a greater probability of being published. As a result, a

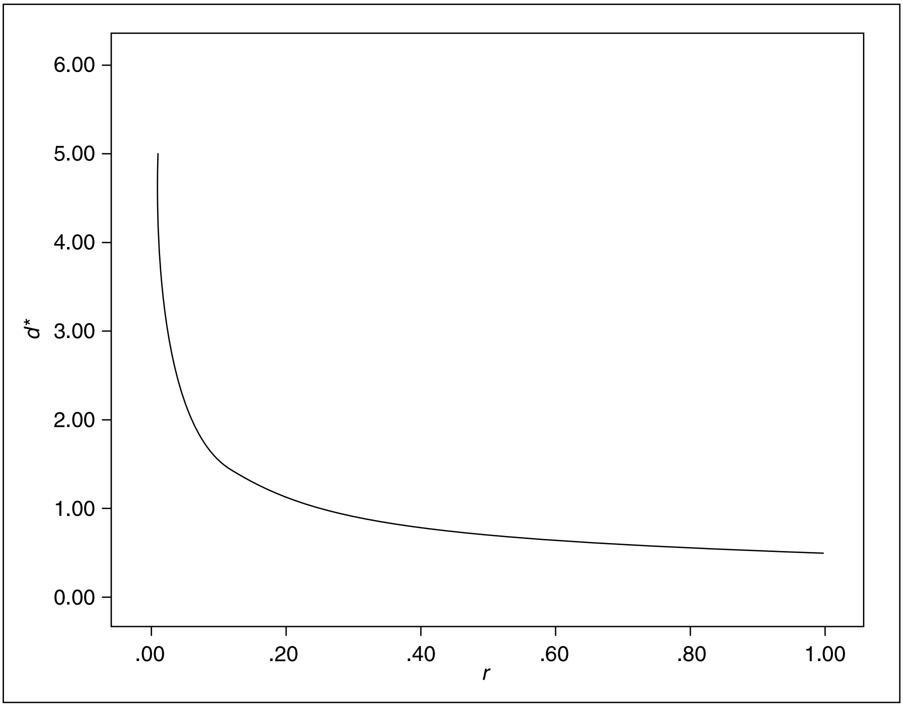


*Figure 11.1* Corrected values for *d* = *.5 for a range of reliabilities.*

sample of published studies can be biased in the direction of reporting large effects. In classical meta-analysis, one way to investigate this issue is to carry out a fail-safe analysis (Rosenthal, 1991). This answers the question of how many unpublished insignificant papers must lie in various researchers' file drawers to render the combined results of the available studies insignificant. If the fail-safe number is high, we assume it is unlikely that the file drawer problem affects our analysis. An alternative approach to the file drawer problem is drawing a *funnel plot*. The funnel plot is a plot of the effect size versus the total sample size (Light & Pillemer, 1984). Macaskill, Walter, and Irwig (2001) recommend using the inverse of the sampling variance instead of the sample size of the studies, because this is a more direct indicator of a study's expected variability; Sterne, Becker, and Egger (2005) suggest using the standard error. These are all indicators of the precision of the studies, and are highly correlated. If the sample of available studies is 'well behaved' the plot should be symmetric and have the shape of a funnel. The outcomes from smaller studies are more variable, but estimate the same underlying population parameter. If large effects are found predominantly in smaller studies, this indicates the possibility of publication bias, and the possibility of many other nonsignificant small studies remaining unpublished in file drawers. In addition to a funnel plot, the effect of study sample size can be investigated directly by including the total sample size of the studies as an explanatory variable in a multilevel meta-analysis. This variable should *not* be related to the outcomes. When instead of sample size the standard error of the effect is included in the model as a predictor, the resulting test is equivalent to the Egger test, a well-known test for funnel asymmetry (Sterne & Egger, 2005).

The example data in Table 11.2 have an entry for the reliability of the outcome measure ($r_{ii}$). These (fictitious) data on the effect of social skills training assume that two different instruments were used to measure the outcome of interest; some studies used one instrument, some studies used another instrument. These instruments, in this example tests for social anxiety in children, differ in their reliability as reported in the test manual. If we use classical psychometric methods to correct for attenuation by unreliability, followed by classical meta-analysis using the random-effects model, the combined effect size is estimated as 0.64 instead of the value of 0.58 found earlier. The parameter variance is estimated as 0.23 instead of the earlier value of 0.17.

If we include the reliability and the sample size as explanatory variables in the regression model, we obtain the results presented in Table 11.4. The predictor variables are centered on their grand mean, to retain the interpretation of the intercept as the 'average outcome'. The first model in Table 11.4 is the empty 'intercept-only' model presented earlier. The second model, which follows equation 11.2, includes the total sample size as a predictor. The third model includes the reliability of the outcome measure. The fourth model includes the duration of the experiment, and the fifth includes all available predictors. Both the univariate and the multivariate analyses show

*Table 11.4* Multilevel meta-analyses on example data

| Model | Intercept-only | + $N_{tot}$ | + reliability | + duration | + all |
|---|---|---|---|---|---|
| Intercept | 0.58 (.11) | 0.58 (.11) | 0.58 (.11) | 0.57 (.08) | 0.58 (.08) |
| $N_{tot}$ | | 0.001 (.01) | | | −0.00 (.01) |
| Reliability | | | .51 (1.40) | | −.55 (1.20) |
| Duration | | | | 0.14 (.04) | 0.15 (.04) |
| $\sigma_u^2$ | 0.14 | 0.16 | 0.16 | 0.04 | 0.05 |
| $p$-value $\chi^2$ deviance | $p < .001$ | $p < .001$ | $p < .001$ | $p = .15$ | $p = .27$ |
| $p$-value $\chi^2$ residuals | $p < .001$ | $p < .001$ | $p < .001$ | $p = .09$ | $p = .06$ |

that only the duration has a significant effect on the study outcomes. Differences in measurement reliability and study size are no major threat to our substantive conclusion about the effect of duration. Since there is no relation between the study size and the reported outcome, the existence of a file drawer problem is unlikely.

The last model that includes all predictor variables simultaneously is instructive. The (nonsignificant) regression coefficient for reliability is negative. This is counterintuitive. This is also in the opposite direction of the regression coefficient in the model (3) with reliability as the only predictor. It is the result of a so-called 'repressor' effect caused by the correlations (from .25 to .33) among the predictor variables. Since in meta-analysis the number of available studies is often small, such effects are likely to occur if we include too many explanatory study-level variables. In the univariate model (11.3), the regression coefficient of reliability is .51. This implies that, if the reliability goes from .75 to .90, the expected outcome increases by (.15 × .51 =) 0.08. This is reasonably close to the correction of 0.06 that results from applying the classical correction for attenuation. However, the large standard error for reliability in model 11.3 suggests that this correction is not needed. Thus, the corrected results using classical methods may well be misleading.

In meta-analysis it is typical to have many study characteristics – and typically many of these are correlated. This leads to substantial multicollinearity, and makes it difficult to determine what effects are important. The approach taken above, to evaluate each effect separately and next look at multiple effects, is a reasonable strategy. The problem of predictor variable selection is a general problem in multiple regression when there are many potential predictors, but it is especially important in meta-analysis because the number of available studies is often small.

## 11.5 MULTIVARIATE META-ANALYSIS

The example in section 11.4 assumes that for each study we have only one effect size, which leads to analysis models with two levels. However, there are several situations that can lead to three-level models. Three-level structures are appropriate if there are multiple studies within the same publication (or multiple studies by the same group of researchers), or if there are multiple effect sizes used in the same study. Such situations lead to a meta-analysis with multiple effect measures, sometimes denoted as a multiple endpoint meta-analysis (Gleser & Olkin, 1994). Three-level structures are also appropriate if the studies investigate the difference between several different treatment groups and one control group. This leads to a collection of effect size measures, which all share the same control group, sometimes denoted as a multiple treatment meta-analysis (Gleser & Olkin, 1994). In both cases, there are dependencies between the effect sizes within studies.

In classical meta-analysis, such dependencies are often ignored by carrying out a series of univariate meta-analyses, or solved by calculating an average effect size across all available outcome measures. For several reasons, this approach is not optimal, and more complex procedures have been proposed to deal with dependent effect sizes (Gleser & Olkin, 1994).

In a multilevel model, we can deal with multiple dependent effect sizes by specifying a multivariate outcome model. Thus, a level is added for the multiple outcome variables, analogous to the multivariate multilevel models discussed in Chapter 10. When some studies do not report on all available outcomes, we have a missing data problem, which is dealt with in the same way as in a standard multivariate multilevel model.

The univariate model for meta-analysis is written as $d_j = \gamma_0 + u_j + e_j$ (see equation 11.4). The corresponding equation for a bivariate random effects meta-analysis is:

$$d_{ij} = \gamma_{0j} + u_{ij} + e_{ij}. \tag{11.11}$$

In equation 11.11, the $j$ is an index that refers to the outcome, and in the bivariate case $j = 1, 2$. The sampling variances of $e_{i1}$ and $e_{i2}$ are assumed to be known, and in addition the covariance between the sampling errors $e_{i1}$ and $e_{i2}$ is assumed to be known. The variances and covariance of the $u_{ij}$ that represent differences between studies are estimated. Thus, replacing the variance terms in the univariate meta-analysis, we have the known covariance matrix $\mathbf{\Omega}_e$ at the second level, and the estimated covariance matrix $\mathbf{\Omega}_u$ at the third level. The lowest level is used only to specify the multivariate structure, following the procedures explained in Chapter 10. Thus, in the bivariate case we have:

$$\mathbf{\Omega}_e = \begin{pmatrix} \sigma^2_{e11} & \sigma_{e1e2} \\ \sigma_{e1e2} & \sigma^2_{e22} \end{pmatrix} \tag{11.12}$$

and:

$$\mathbf{\Omega}_u = \begin{pmatrix} \sigma^2_{u11} & \sigma_{u1u2} \\ \sigma_{u1u2} & \sigma^2_{u22} \end{pmatrix}. \tag{11.13}$$

The covariance between $e_1$ and $e_2$ can also be written as $\sigma_{e1e2} = \sigma_{e1}\,\sigma_{e2}\rho_W$, where $\rho_W$ is the known within-study correlation. Generally, $\rho_W$ is estimated by the correlation between the outcome variables in the control group or the pooled correlation across the control and experimental group (Gleser & Olkin, 1994). From the estimated matrix $\mathbf{\Omega}_u$ we can calculate the between-studies correlation $\rho_B$, using $\rho_B = \sigma_{u1u2}/\sigma_{u1}\,\sigma_{u2}$.

Table 11.5 presents the formula for the sampling covariance of some effect measures (see Raudenbush & Bryk, 2002). The covariance between two correlations is a complicated expression discussed in detail by Steiger (1980) and presented in an accessible manner by Becker (2007).

*Table 11.5* Some effect measures, their transformation, and sampling covariance

| Measure | Estimator | Transformation | Sampling covariance |
|---|---|---|---|
| Mean | $\bar{x}$ | – | $s^2/n$ |
| Diff. 2 means | $g = (\bar{y}_E - \bar{y}_C)$ | $d = (1 - 3/(4N - 9))g$ | $\rho_w(n_E + n_C)/(n_E\, n_C) + \rho_w^2 d_1 d_2/(2(n_E + n_C))$ |
| Stand. dev. | $S$ | $s^* = \ln(s) + 1/(2df)$ | $\rho_w^2/(2df)$ |
| Correlation | $r$ | – | $\sigma(r_{st}, r_{uv}) = [0.5\, r_{st}, r_{uv}\, (r_{su}^2 + r_{sv}^2 + r_{tu}^2 + r_{tv}^2) + r_{su}r_{tv} + r_{sv}r_{tu} - (r_{st}r_{su}r_{sv} + r_{ts}r_{tu}r_{tv} + r_{us}r_{uv}r_{ut} + r_{vs}r_{vtvu}r)]/n$ |
| Proportion | $P$ | $\text{logit} = \ln\,(p/1 - p)$ | $1/[(np\,(1 - p_1))\,(np\,(1 - p_2))]$ |

Currently HLM is the only software that directly inputs the vector of effect sizes and the sampling (co)variances, although other software can be used with special command setups. These and other software issues are discussed in the appendix to this chapter.

A serious limitation for multivariate meta-analysis is that the required information on the correlations between the outcome variables is often not available in the publications. Some approximations may be available. For instance, if standardized tests are used, the test manual generally provides information on the correlations between subtests. If a subset of the studies reports the relevant correlations, they can be meta-analyzed in a preliminary step, to obtain a global estimate of the correlation between the outcomes. Riley, Thompson, and Abrams (2008) suggest setting the within-study covariance equal to the covariance between the effect measures. Alternatively, researchers can conduct a 'sensitivity analysis' in which a range of plausible values for the correlation between outcome measures can be used to determine the likely effect of this issue on substantive interpretations. Cohen (1988) suggested using the value of .10 for a small correlation, .30 for a medium correlation, and .50 for a large correlation. Taking these suggestions as a starting point, a sensitivity analysis using the values .00, .10, .30, and .50 appears reasonable.

To illustrate the proceedings we analyze a bivariate meta-analytic data set discussed by Nam, Mengersen, and Garthwaite (2003). The data are from a set of studies that investigate the relationship between children's environmental exposure to smoking (ETS) and the child health outcomes of asthma and lower respiratory disease (LRD). Table 11.6 lists the logged odds ratio (LOR) for asthma and LRD, and their standard errors (SE in Table 11.6). Study-level variables are the average age of subjects, publication year, smoking (0 = parents, 1 = other in household), and covariate adjustment used (0 = not, 1 = yes).

*Table 11.6* Selected data from studies reporting odds ratios for asthma and LRD

| ID | Size | Age | Year | Smoke | LOR asthma | SE LOR asthma | LOR LRD | SE LOR LRD |
|---|---|---|---|---|---|---|---|---|
| 3 | 1285 | 1.1 | 1987 | 0 | | | .39 | .27 |
| 4 | 470 | 9.0 | 1994 | 0 | .04 | .20 | | |
| 6 | 1077 | 6.7 | 1995 | 0 | | | .35 | .15 |
| 8 | 550 | 1.7 | 1995 | 0 | .61 | .18 | | |
| 10 | 850 | 9.4 | 1996 | 0 | | | .25 | .23 |
| 11 | 892 | 10.9 | 1996 | 0 | | | −.02 | .22 |
| 14 | 1232 | 9.5 | 1996 | 0 | | | −.09 | .27 |
| 16 | 3048 | 1.0 | 1997 | 0 | .42 | .12 | | |
| 17 | 2216 | 8.6 | 1997 | 1 | | | −.27 | .15 |
| 19 | 535 | 6.5 | 1995 | 0 | | | .30 | .60 |
| 20 | 5953 | 0.5 | 1987 | 0 | .99 | .20 | | |
| 22 | 159 | 1.0 | 1986 | 0 | .69 | .24 | | |
| 24 | 9670 | 5.0 | 1989 | 0 | .05 | .09 | −.04 | .12 |
| 25 | 318 | 8.2 | 1995 | 0 | | | .34 | .36 |
| 26 | 343 | 9.5 | 1995 | 0 | .85 | .28 | | |
| 28 | 11534 | 9.5 | 1996 | 1 | .12 | .06 | −.02 | .11 |
| 29 | 10106 | 7.5 | 1984 | 0 | −.02 | .06 | | |
| 32 | 443 | 3.8 | 1992 | 0 | 1.93 | .46 | | |
| 36 | 12727 | 2.5 | 1987 | 0 | .38 | .10 | | |
| 37 | 257 | 7.5 | 1989 | 1 | .99 | .37 | | |
| 38 | 511 | 2.5 | 1995 | 0 | | | .52 | .17 |
| 40 | 253 | 1.0 | 1995 | 0 | −.33 | .51 | | |
| 43 | 1503 | 13.0 | 1997 | 0 | .12 | .14 | | |
| 44 | 7677 | 14.0 | 1995 | 0 | .14 | .02 | .16 | .06 |
| 49 | 1001 | 7.0 | 1988 | 1 | .04 | .19 | | |
| 50 | 961 | 13.5 | 1995 | 0 | .02 | .13 | | |

*Table 11.6* Continued

| ID | Size | Age | Year | Smoke | LOR asthma | SE LOR asthma | LOR LRD | SE LOR LRD |
|---|---|---|---|---|---|---|---|---|
| 51 | 1138 | 9.0 | 1983 | 0 | .01 | .11 | | |
| 52 | 5412 | 9.2 | 1997 | 1 | .22 | .09 | .30 | .09 |
| 54 | 925 | 7.0 | 1997 | 1 | .53 | .32 | .47 | .28 |
| 56 | 2554 | 9.5 | 1995 | 1 | | | −.56 | .33 |
| 57 | 4665 | 13.5 | 1997 | 1 | .04 | .07 | | |
| 59 | 1501 | 9.5 | 1991 | 1 | .18 | .15 | .10 | .11 |
| 61 | 1143 | 1.0 | 1981 | 0 | .04 | .01 | | |
| 63 | 483 | 1.0 | 1989 | 0 | .48 | .23 | | |
| 65 | 8154 | 4.6 | 1995 | 0 | .20 | .04 | | |
| 69 | 2077 | 0.5 | 1974 | 0 | .50 | .21 | | |
| 71 | 8585 | 7.0 | 1993 | 0 | | | .19 | .04 |
| 75 | 301 | 1.0 | 1988 | 1 | .69 | .35 | | |
| 76 | 153 | 10.9 | 1989 | 0 | .92 | .39 | | |
| 78 | 3072 | 8.5 | 1982 | 0 | | | .40 | .16 |
| 79 | 4331 | 2.5 | 1990 | 0 | | | .74 | .23 |
| 80 | 1215 | 2.0 | 1993 | 1 | | | .79 | .21 |
| 81 | 708 | 10.5 | 1994 | 1 | | | .34 | .19 |
| 82 | 620 | 10.0 | 1995 | 0 | | | 1.03 | .39 |
| 83 | 126 | 9.0 | 1991 | 0 | | | .94 | .37 |
| 84 | 228 | 8.5 | 1992 | 1 | | | .69 | .29 |
| 85 | 914 | 3.5 | 1993 | 0 | | | .36 | .18 |
| 88 | 16562 | 8.0 | 1991 | 0 | .30 | .30 | | |
| 89 | 15712 | 5.0 | 1995 | 0 | .29 | .19 | | |
| 93 | 3482 | 8.4 | 1986 | 0 | .22 | .08 | .14 | .13 |
| 105 | 192 | 0.8 | 1992 | 0 | 1.19 | .50 | | |
| 109 | 726 | 16.3 | 1996 | 0 | 1.47 | .47 | | |
| 111 | 372 | 7.0 | 1980 | 0 | 1.77 | 1.03 | | |
| 113 | 684 | 9.0 | 1988 | 1 | .22 | .11 | −.07 | .04 |
| 114 | 207 | 9.3 | 1992 | 1 | | | .36 | .33 |
| 122 | 700 | 7.9 | 1992 | 0 | | | .35 | .19 |
| 601 | 9653 | 9.0 | 1997 | 1 | | | .44 | .14 |
| 603 | 3769 | 9.0 | 1997 | 1 | | | .10 | .23 |
| 603 | 2540 | 9.0 | 1997 | 1 | | | −.36 | .43 |

There are two effect sizes, the logged odds ratios for asthma and for lower respiratory disease (LRD). Only a few studies report both. There is no information on the correlation between LOR asthma and LRD within the studies, but at the study level the correlation is .80. However, this is an ecological correlation, which confounds within-study and between-study effects. The analysis choices are to model the within-study variances only, setting the covariance to zero, or to assume a common value for the within-study variance (e.g., $r = .30$, a medium size correlation). As a first approximation, we set the covariance to zero. To analyze these data, the data file in Table 11.6 must be restructured into a 'long' or 'stacked' file format (see Chapter 10), where the outcomes for asthma and LRD become conditions $i$ nested within studies $j$. Since many studies report only one outcome, these studies have only one condition. The intercept-only model for these bivariate data is:

$$LOR_{ij} = \beta_{0j}Asthma + \beta_{1j}LRD + e(A)_{ij} + e(L)_{ij}. \tag{11.14}$$

In equation 11.14, the variables *Asthma* and *LRD* are dummy variables referring to the asthma and LRD outcomes. The error terms $e(A)_{ij}$ and $e(L)_{ij}$ are also specific for each outcome, and are assumed to be uncorrelated.

Table 11.7 presents the results of a series of models for the exposure to smoking data. The intercept-only model shows a clear effect; children who are exposed to environmental smoking face increased odds of having asthma or LRD. The variances are significant by the Wald test. Both the univariate deviance difference test (removing the study-level variances individually) and the multivariate deviance difference test (removing the study-level variances simultaneously) confirm this. The effect on LRD

*Table 11.7* Bivariate meta-analysis for exposure to smoking, covariance constrained to zero

| Model | Intercept only | Equality | + Age (centered) |
|---|---|---|---|
| **Fixed part** | Coeff. (s.e.) | Coeff. (s.e.) | Coeff. (s.e.) |
| Interc. Asthma | 0.32 (.04) | 0.29 (.04) | 0.29 (.03) |
| Interc. LRD | 0.27 (.05) | 0.29 (.04) | 0.29 (.03) |
| Age | | | –0.03 (.006) |
| **Random part** | | | |
| Var (Asthma) | 0.06 (.02) | 0.08 (.03) | 0.06 (.02) |
| Var (LRD) | 0.07 (.02) | 0.05 (.02) | 0.03 (.01) |
| Covar (AL) | 0.06 (.02) | 0.06 (.02) | 0.05 (.01) |
| Deviance | 44.2 | 44.7 | 32.4 |

appears somewhat stronger than the effect on asthma. In a multivariate analysis this difference can be tested by a Wald test, or by constraining these regression coefficients to be equal. The Wald test for this equality constraint produces a chi-square of 1.232, with $df = 1$ and $p = .27$ clearly not significant. The column marked 'equality' in Table 11.7 reports the estimates when an equality constraint is imposed on the regression coefficients for asthma and LRD. The deviance difference test can not be used here to test the difference, since the estimates use RML. The last column adds the variable Age, which is the only significant study-level predictor. Older children show a smaller effect of exposure to smoking. Age is entered as a single predictor, not as an interaction with the asthma or LRD dummy. This assumes that the effect of age on asthma and LRD is the same. Exploratory analysis shows indeed very similar regression coefficients when asthma and LRD are modeled separately in a bivariate meta-analysis, and the Wald equality test on the regression coefficients is not significant ($p = .45$).

The results reported in Table 11.7 are estimates where the common correlation between the outcomes is constrained to be zero. This causes some bias; Riley et al. (2008) report a simulation that shows that this leads to an upward bias in the study-level variances, which in turn leads to some bias in the fixed effects and increased standard errors for the fixed effects. They recommend either imputing a reasonable value for $r$, or estimating only one covariance parameter that confounds the within and between study-level covariance. They report simulations that show that the latter strategy is quite successful. In our case, we carry out a sensitivity analysis, where several plausible values are specified for the covariance between the error terms $e(A)_{ij}$ and $e(L)_{ij}$. Since $e(A)_{ij}$ and $e(L)_{ij}$ are standardized to have a variance equal to one, the covariance is equal to the correlation. Table 11.8 shows the results when the common

*Table 11.8* Results of bivariate meta-analysis for exposure to smoking, for three values of covariance between outcomes

| Covariance = | .10 | .30 | .50 | common |
|---|---|---|---|---|
| **Fixed part** | Coeff. (s.e.) | Coeff. (s.e.) | Coeff. (s.e.) | Coeff. (s.e.) |
| Interc. Asthma | 0.29 (.03) | 0.29 (.03) | 0.29 (.03) | 0.29 (.03) |
| Interc. LRD | 0.29 (.03) | 0.29 (.03) | 0.29 (.03) | 0.29 (.03) |
| Age | –.03 (.006) | –.03 (.006) | –.03 (.006) | –.03 (.006) |
| **Random part** | | | | |
| Var (Asthma) | 0.03 (.01) | 0.03 (.01) | 0.03 (.01) | 0.03 (.01) |
| Var (LRD) | 0.06 (.02) | 0.06 (.02) | 0.07 (.02) | 0.06 (.02) |
| Covar (AL) | 0.05 (.01) | 0.05 (.01) | 0.05 (.01) | 0.05 (.01) |
| Deviance | 32.6 | 32.3 | 31.6 | 32.7 |

correlation is constrained to .1, .3, and .5 (Cohen's suggestions for a small, medium, and large correlation), and when a single common covariance is estimated for the between-study and the within-study part. Two conclusions are evident. First, the estimated effect of passive smoking on asthma and LRD is similar to the results reported in Table 11.7, and second, all results are remarkably similar. It should be noted that in these meta-analytic data the variation between studies is small. Riley et al. (2008) report simulations that show that with larger between-study variation the differences are larger. Still, they also report that the estimates for the fixed effects are not affected much by the different specifications for the within-study correlation.

Multivariate meta-analysis is especially useful if most studies do not report on all possible outcomes. A series of univariate meta-analyses on such data assumes that the missing outcome variables are missing completely at random. A multivariate meta-analysis assumes that the missing outcomes are missing at random, a less strict assumption. In addition, a multivariate meta-analysis allows testing of the equality of effect sizes and regression coefficients, as exemplified in the bivariate exposure to smoking example.

For details on multivariate multilevel meta-analysis see Kalaian and Raudenbush (1996), Normand (1999), van Houwelingen, Arends, and Stijnen (2002), and Kalaian and Kasim (2008). An example of a multivariate multilevel meta-analysis is discussed by Berkey et al. (Berkey, Hoaglin, Antczak-Bouckoms, Mosteller, & Colditz, 1998).

Another interesting extension of multilevel meta-analysis arises when we have access to the raw data for at least some of the studies. This situation leads to a multilevel model that combines both sufficient statistics, as in standard meta-analysis, and raw data to estimate a single effect size parameter. Higgins et al. (Higgins, Whitehead, Turner, Omar, & Thompson, 2001) describe the general framework for this hybrid meta-analysis, and discuss classical and Bayesian analysis methods. Examples of such hybrid meta-analyses are the studies by Goldstein et al. (Goldstein, Yang, Omar, Turner, & Thompson, 2000) and Turner et al. (Turner, Omar, Yang, Goldstein, & Thompson, 2000).

## 11.6 STATISTICAL AND SOFTWARE ISSUES

The program HLM (Raudenbush et al., 2004) has a built-in provision for meta-analysis, which is restricted to two levels. If we need three levels, we can use the standard HLM software, using an adapted program setup. Other software can be used, provided it is possible to put restrictions on the random part. MLwiN (Rasbash et al., 2000) and Proc Mixed in SAS (Littell et al., 1996) have this capacity, and can therefore be used for meta-analysis, again with an adapted setup. Ways of tweaking HLM and MLwiN for meta-analysis are discussed in the appendix to this chapter.

There are some minor differences between the programs. HLM uses by default an estimator based on restricted maximum likelihood (RML), while MLwiN by default uses full maximum likelihood (FML, called IGLS in MLwiN). Since RML is theoretically better, especially in situations where we have small samples and are interested in the variances, for meta-analysis we should prefer RML (called RIGLS in MLwiN). If in a specific case the difference between RML and FML is small, we can choose FML because it allows testing regression coefficients using the deviance difference test. The results reported in this chapter were all estimated using RML.

An important difference between HLM and other multilevel analysis software is the test used to assess the significance of the variances. HLM by default uses the variance test based on a chi-square test of the residuals (Raudenbush & Bryk, 2002; see Chapter 3 of this book). MLwiN estimates a standard error for each variance, which can be used for a $Z$-test of the variance. In meta-analysis applications, this $Z$-test is problematic. First, it is based on the assumption of normality, and variances have a chi-square distribution. Second, it is a large-sample test, and with small sample sizes and small variances the $Z$-test is inaccurate. In meta-analysis the sample size is the number of studies that are located, and it is quite common to have at most 20 studies. An additional advantage of the chi-square test on the residuals is that for the empty model this test is equivalent to the chi-square variance test in classical meta-analysis (Hedges & Olkin, 1985). The variance tests reported in this chapter use both the deviance difference test and the chi-square test on the residuals. MLwiN does not offer this test, but it can be produced using the MLwiN macro language.

It should be noted that the standard errors that are used to test the significance of the regression coefficients and to establish confidence intervals are also asymptotic. With the small samples common in meta-analysis, they can lead to confidence intervals that are too small, and $p$-values that are spuriously low (Brockwell & Gordon, 2001). It appears prudent not to use the standard normal distribution, but the Student $t$-distribution with degrees of freedom equal to $k - p - 1$, where $k$ is the number of studies and $p$ the number of study-level explanatory variables in the model. In HLM this is the standard test for the regression coefficients. In simulations by Berkey et al. (1998) this provided correct $p$-values. Brockwell and Gordon (2001) recommend profile likelihood methods and bootstrapping. These are treated in Chapter 13 in this book.

For estimating complex models, Bayesian procedures are promising and are coming into use (see Sutton et al., 2000). These use computer-intensive methods such as Markov chain Monte Carlo (MCMC) to estimate the parameters and their sampling distributions. These methods are attractive for meta-analysis (DuMouchel, 1989, 1994; Smith, Spiegelhalter, & Thomas, 1995) because they are less sensitive to the problems that arise when we model small variances in small samples. Bayesian models are treated in Chapter 13 in this book. Bayesian modeling starts with the specification of a prior distribution that reflects a priori beliefs about the distribution of the parameters. In principle, this is an elegant method to investigate the effect of publication bias. An example of such an analysis is Biggerstaff, Tweedy, and Mengersen (1994). Although the software MLwiN includes Bayesian methods, at present these cannot analyze meta-analytic models, and more complicated software is needed, such as the general Bayesian modeling program BUGS (Spiegelhalter, 1994). Nam et al. (2003) discuss several Bayesian meta-analysis models, using the exposure to environmental smoking data as their example.

## APPENDIX

### Software implementation: General

Multilevel meta-analysis requires software that allows imposing constraints on the random part. If that is possible, there are two different but functionally equivalent methods to specify the known variance. The first method is to add the standard error of the outcome to the model as a predictor. This predictor should be in the random part at the first level only, and not in the fixed part or elsewhere in the random part. Next, the variance for this predictor is constrained to be equal to 1. The second method is to use the inverse of the sampling variance (the square of the standard error) as a weight at the first level. The weights must be raw weights, not normalized. Next, the variance at the lowest level is constrained to be 1. The weight method is the most general, for instance, Cheung (2008) uses it to carry out a meta-analysis in the structural equation software Mplus.

In HLM and MLwiN both methods are feasible. To apply multilevel models in meta-analysis in other software, such as SAS Proc Mixed, the software must have options to set up a model using constraints as specified for MLwiN or for HLM. This means that it must be possible to have a complex lower-level variance structure, as in MLwiN, or to constrain the lowest-level variance to 1 and to add a weight variable, as in HLM. So far, public domain software for multilevel analysis does not offer these options. For univariate meta-analysis, David Wilson (Lipsey & Wilson, 2001) has written macros for SAS, SPSS, and STATA that carry out both ML and RML based meta-regression, equivalent to the multilevel approach.

### Software implementation: HLM

The simplest program for multilevel meta-analysis is the meta-analysis module that is part of the software HLM. HLM expects for each study a row of data containing the study ID, an effect measure, and its sampling variance, followed by the explanatory variables. This software can also carry out a multivariate meta-analysis, but does use listwise deletion on the studies that have missing outcome variables.

HLM can also be used for meta-analysis using the weight method described earlier. In HLM, the weight that is supplied is the sampling variance itself, and the estimation specification must include a specification that the weights are sampling variance weights. From that point, the analysis proceeds as a standard HLM analysis.

If we need more than two levels in HLM, we must use HLM with the inverse sampling variance weight method. Multivariate multilevel meta-analysis is done using a separate level for the outcome variables, as illustrated earlier.

HLM can be tweaked to analyze a multivariate meta-analysis using the sampling variance weighting method. This specifies the within-study variance correctly, but assumes uncorrelated outcomes. Kalaian and Raudenbush (1996) describe a method to transform the multivariate outcome matrices to make the outcomes orthogonal, without disturbing the parameter estimates for the fixed part.

**Software implementation: MLwiN**

Using MLwiN is more complicated. The data structure is analogous to HLM: we need a study ID, the effect size, its standard error (the square root of the sampling variance), the regression constant (HLM includes this automatically), and the explanatory variables. To set up the analysis, we distinguish two levels: the outcomes are the first level and the studies are the second. Usually we have one outcome per study, so there is no real nesting. The predictor ‘standard error’ is included only in the random part on level 1, with a coefficient fixed at 1. The regression constant is included in the fixed part and in the random part at level 2. Explanatory variables are included in the fixed part only. MLwiN does not produce the chi-square test on the variances. The formula for the chi-square test is $\chi^2 = \sum\left(\left(d_j - \hat{d}_j\right) / s.e.\left(d_j\right)\right)^2$, which is the sum of the squared residuals divided by their standard errors. The degrees of freedom are given by $df = J - q - 1$, where $J$ is the number of studies, and $q$ the number of explanatory variables in the model. Assuming that the outcomes are denoted by ‘*d*’, and the standard errors of the *d*s by ‘*sed*’, the sequence of MLwiN commands for computing the chi-square is:

- PRED C50 (produce predicted values)
- CALC C50=((‘*d*’ – C50)/‘*sed*’)∧2 (produce squared residuals/s.e.)
- SUM C50 to B1 (sum to box B1)
- CPRO B1 *df* (calculate *p*-value).

This code assumes that the spreadsheet column C50 is unused.

In a multivariate meta-analysis the setup is more complicated. There are multiple outcomes, with each outcome indicated by a 0–1 dummy variable that indicates that outcome. There is no intercept in the model, so for *p* outcomes all *p* dummy variables can be used. This structure can be generated automatically in MLwiN using the multiple response option. There are as many standard error variables as there are outcome variables. To accomplish this, the standard error is multiplied by each of the dummy variables, which produces a column with standard errors for that specific outcome, and zeros elsewhere. MLwiN can do this automatically when the multiple response option is activated. The standard error variables are all entered only in the random part at the lowest level, with their variances constrained to 1. The covariances are constrained to zero or to their common known or imputed value. In a multivariate meta-analysis the residuals chi-square must be calculated for each outcome separately. For example, for the bivariate meta-analysis on the log odds ratio (LOR) in the exposure to smoking example, the sequence of MLwiN commands is:

- PRED C50
- CALC C50 = ((‘LOR’ – C50)/ ‘SE_LOR’)^2
- CALC C51 = C50*‘ASTHMADUMMY’
- SUM C51 TO B1
- CPRO B1 *df*
- CALC C52 = C50*‘LRDDUMMY’
- SUM C52 TO B2
- CPRO B2 *df*

Multiplication of the values in C50 by the asthma dummy and the LRD dummy creates columns C51 and C52 that contain only values for asthma and LRD respectively. These are summed separately. In the exposure to smoking example, this produces a value for chi-square of 148.6 for asthma and 1041.0 for LRD. For asthma we have *df* = 32 – 1 = 31 and for LRD we have *df* = 35 – 1 = 34. Both chi-square values are highly significant.

MLwiN can also use the inverse sampling variance weight method. In a multivariate meta-analysis, the standard errors as variables method is preferred, because then we can specify a correlation between the outcomes. In the inverse sampling weight method, this correlation is assumed to be zero.

# 12

# Sample Sizes and Power Analysis in Multilevel Regression

Questions about sample size tend to focus on two topics: what sample size is sufficient to apply a specific statistical method, and what sample size is needed to obtain a specific power. In multilevel analysis, these problems are made more difficult because there are sample size issues at more than one level, and because the model includes a fixed and a random part, and typically the fixed part can be estimated with more precision than the random part. This chapter reviews both the sample size and the power issues.

## 12.1 SAMPLE SIZE AND ACCURACY OF ESTIMATES

The maximum likelihood estimation methods used commonly in multilevel analysis are asymptotic, which translates to the assumption that the sample size is large. This arouses questions about the accuracy of the various estimation methods with relatively small sample sizes. Most research on this problem uses simulation methods, and investigates the accuracy of the fixed and random parameters with small sample sizes at either the individual or the group level. Comparatively less research investigates the accuracy of the standard errors used to test specific model parameters.

### 12.1.1 Accuracy of fixed parameters and their standard errors

The estimates for the regression coefficients are generally unbiased for ordinary least squares (OLS), generalized least squares (GLS), and maximum likelihood (ML) estimation (Maas & Hox, 2004a, 2004b; van der Leeden & Busing, 1994; van der Leeden et al., 2008). OLS estimates are less efficient because they often have a larger sampling variance. Kreft (1996) reports that OLS estimates are about 90% efficient.

As illustrated in Chapter 2, the OLS-based standard errors are severely biased downward. The asymptotic Wald tests, used in most multilevel software to test fixed effects, assume large samples. A large simulation by Maas and Hox (2004a) finds that the standard errors for the fixed parameters are slightly biased downward if the number of groups is less than 50. With 30 groups, they report an operative alpha level of 6.4% while the nominal significance level is 5%. Similarly, simulations by van der

Leeden and Busing (1994) and van der Leeden et al. (2008) suggest that when assumptions of normality and large samples are not met, the standard errors have a small downward bias. GLS estimates of fixed parameters and their standard errors are less accurate than ML estimates.

The power of the Wald test for the significance of the individual-level regression coefficients depends on the total sample size. The power of tests of higher-level effects and cross-level interactions depends more strongly on the number of groups than on the total sample size. Both simulations (Mok, 1995; van der Leeden & Busing, 1994) and analytic work (Cohen, 1998; Raudenbush & Liu, 2000; Snijders & Bosker, 1993) suggest a trade-off between sample sizes at different levels. For accuracy and high power a large number of groups appears more important than a large number of individuals per group.

### 12.1.2 Accuracy of random parameters and their standard errors

Estimates of the residual error at the lowest level are generally very accurate. The group-level variance components are sometimes underestimated. Simulations by Busing (1993) and van der Leeden and Busing (1994) show that GLS variance estimates are less accurate than ML estimates. The same simulations also show that for accurate group-level variance estimates many groups (more than 100) are needed (see Afshartous, 1995). However, using later versions of the software MLn, Browne and Draper (2000) show that with as few as 6–12 groups, restricted ML (RML) estimation can provide reasonable variance estimates. With 48 groups, full ML (FML) estimation also produces good variance estimates. Maas and Hox (2004a) report that with as few as 30 groups, RML estimation produces accurate variance estimates. When the number of groups is around 10, the variance estimates are much too small.

The asymptotic Wald test for the variance components implies the unrealistic assumption that they are normally distributed. For this reason, other approaches have been advocated, including estimating the standard error for sigma (the square root of the variance, Longford, 1993), and using the likelihood ratio test. Bryk and Raudenbush (1992) advocate a chi-square test based on the OLS residuals. The literature contains no comparisons between all these methods. Simulations by van der Leeden et al. (1997) show that, especially with small numbers of small groups, estimates of the standard errors used for the Wald test are often too low, with RML again more accurate than FML. Symmetric confidence intervals around the estimated value also do not perform well. Browne and Draper (2000) and Maas and Hox (2004a) report similar results. Typically, with 24–30 groups, the operating alpha level was almost 9%, and with 48–50 groups about 8%. In the simulations by Maas and Hox (2004a), with 100 groups the operating alpha level was 6%, which is close to the nominal 5%. Chapter 13 of this book treats some alternatives to the asymptotic Wald tests, which may be preferable when small variance components are tested or when the number of groups is less than 50.

### 12.1.3 Accuracy and sample size at different levels

It is clear that with increasing sample sizes at all levels, estimates and their standard errors become more accurate. Kreft (1996) suggests a rule of thumb, which she calls the '30/30 rule'. To be on the safe side, researchers should strive for a sample of at least 30 groups with at least 30 individuals per group. From the various simulations, this seems sound advice if the interest is mostly in the fixed parameters. For certain applications, one may modify this rule of thumb. Specifically, if there is strong interest in cross-level interactions, the number of groups should be larger, which leads to a 50/20 rule: about 50 groups with about 20 individuals per group. If there is strong interest in the random part, the variance and covariance components and their standard errors, the number of groups should be considerably larger, which leads to a 100/10 rule: about 100 groups with at least about 10 individuals per group.

These rules of thumb take into account that there are costs attached to data collection, so if the number of groups is increased, the number of individuals per group decreases. In some cases, this may not be a realistic reflection of costs. For instance, in school research an extra cost will be incurred when an extra class is included. Testing only part of the class instead of all pupils will usually not make much difference in the data collection cost. Given a limited budget, an optimal design should reflect the various costs of data collection. Snijders and Bosker (1993), Cohen (1998), Raudenbush and Liu (2000), and Moerbeek, van Breukelen, and Berger (2000) all discuss the problem of choosing sample sizes at two levels while considering costs. Moerbeek, van Breukelen, and Berger (2001) discuss the problem of optimal design for multilevel logistic models. Essentially, optimal design is a question of balancing statistical power against data collection costs. Data collection costs depend on the details of the data collection method (see Groves, 1989). The problem of estimating power in multilevel designs is treated later in this chapter.

### 12.1.4 Accuracy and sample size with proportions and dichotomous data

Multilevel analysis of proportions generally uses generalized linear models with a logit link (see Chapter 6), which gives us the model:

$$\pi_{ij} = \text{logistic}(\gamma_{00} + \gamma_{10}X_{ij} + u_{0j}) \tag{12.1}$$

The observed proportions $P_{ij}$ are assumed to have a binomial distribution with known variance:

$$\mathrm{var}(P_{ij}) = (\pi_{ij}\,(1 - \pi_{ij}))/n_{ij} \qquad (12.2)$$

The $\pi_{ij}$ are estimated by prediction from the current model. If the variance term is not constrained to one, but estimated, we can model over- and underdispersion. If the extrabinomial variation is significantly different from one, this is usually interpreted as an indication that the model is misspecified, for instance by leaving out relevant levels, interactions among predictors, or in time series data by not allowing autocorrelation in the error structure.

Most programs rely on a Taylor expansion to linearize the model. The program MLwiN uses a first order Taylor expansion and marginal (quasi) likelihood (MQL1: $P_{ij}$ predicted by fixed part only) as a default. MLwiN can also use a second order expansion and predictive or penalized (quasi) likelihood (PQL2: $P_{ij}$ predicted by both fixed and random part), while HLM uses first order expansion and predictive or penalized (quasi) likelihood (PQL1).

Simulations by Rodriguez and Goldman (1995, 2001) show that the marginal quasi-likelihood method with first order Taylor expansion (MQL1) underestimates both the regression coefficients and the variance components, in some cases very severely. Goldstein and Rasbash (1996) compare MQL1 and PQL2 by simulating data according to the worst performing data set of Rodriguez and Goldman. This is a three-level data set, with 161 communities that contain in total 1558 women who reported 2449 births. Therefore, each community has on average 9.7 women, who on average report on 1.6 births. In Goldstein and Rasbash's simulation, the means of the MQL1 estimates for the fixed effects, from 200 simulation runs, were underestimated by about 25%. The means of the MQL1 estimates for the random effects were underestimated by as much as 88%. Moreover, 54% of the level 2 variances were estimated as zero, while the population value is 1. For the same 200 simulated data sets, the means of the PQL2 estimates for the fixed effects underestimated the population value by at most 3%, and for the random effects by at most 20%. None of the PQL2 variance estimates was estimated as zero.

Browne and Draper (2000) also report a simulation study based on the structure of the Rodriguez and Goldman data. In their simulation, the MQL1 method had an abysmal performance. The PQL2 method fares somewhat better: the regression coefficients are close, but the actual coverage of the 95% confidence interval is close to 95% only for the lowest-level predictor; for the predictors at the woman and the community level the actual coverage is about 90%. The variances are still not estimated very accurately: the PQL2 method underestimates the woman-level variance by 11% and the community-level variance by 43%, and the actual coverage of the 95% confidence interval for the variance estimates is 78% for the woman level and 27% for the community level.

If anything, the analysis of proportions and binomial data requires larger samples than the analysis of normally distributed data. The Rodriguez and Goldman data set is extreme, because the data are dichotomous, the variance components are large, and the sample size at the lowest level is very small. Consequently, the estimated proportions at the lowest level are very inaccurate. In less extreme cases, it appears that penalized quasi likelihood with second order Taylor expansion is usually sufficiently accurate for the regression coefficients and in many cases good enough for the random parameters. A review of the available literature shows that PQL-based estimates and tests for the regression coefficients are accurate with samples of modest sizes, but estimates and tests of the variances are not. However, with some data sets the PQL2 algorithm breaks down, and the MLwiN manual recommends beginning with the simpler MQL1 approach to obtain good starting values for the more complicated PQL2 approach. Using maximum likelihood estimation with numerical integration of the likelihood provides generally better estimates, but does not compensate for the fact that binary data contain less information than normal data, and therefore need larger sample sizes for accurate estimation. Moineddin, Matheson, and Glazier (2007) report a simulation study on multilevel logistic regression using numerical integration (SAS NLMixed procedure). They find that multilevel logistic models require larger sample sizes than models for normal data, and that the sample size requirements increase when the modeled proportions are close to zero or one. They recommend having at least 50 groups with a group size of 50.

For problematic data structures, such as proportions very close to 0 or 1 and small sample sizes, bootstrapping approaches and Bayesian estimation using Gibbs sampling offer improvements. These are described in Chapter 13.

## 12.2 ESTIMATING POWER IN MULTILEVEL REGRESSION DESIGNS

Statistical testing controls the risk of erroneously rejecting the null hypothesis ($H_0$) or committing a type I error by setting a significance level $\alpha$. The alpha level is the maximum probability tolerated for falsely rejecting the null-hypothesis. By convention, it is usually set as $\alpha = .05$, and less often to the more rigorous $\alpha = .01$. Sometimes, as in explorative analyses, the more lenient $\alpha = .10$ is chosen.

When the null-hypothesis is false, it should be rejected in favor of the alternative hypothesis $H_A$, which states that a certain effect exists. Failure to reject the null-hypothesis in this case implies another error, denoted by $\beta$ or type II error. The probability of committing this error is as a rule discussed in terms of the *power* of the statistical test: that is, the probability of rejecting the null-hypothesis when it is in fact not true. Power increases when $\alpha$ is set to a higher level, and with larger samples or larger effect sizes. In the so-called Newman-Pearson approach to hypothesis testing (Barnett, 1999), a specific value for the alternative hypothesis $H_A$ is stipulated, and both $\alpha$ and $\beta$ are chosen to balance the relative costs of committing a type I or a type II error. In the absence of clear conceptions of these costs, Cohen (1988, 1992) recommends using a power of .80 (corresponding to $\beta = 0.20$) as a conventional value for a high power, because this is an adequate power level, which still keeps the sample size requirements within acceptable limits. A power level of .50 is considered moderate.

The power of a statistical test is a function of the significance level, the sample size, and the population effect size. For decisions about the research design and sample size to be employed, it is useful to estimate the sample size that is needed to achieve a specific power for a given $\alpha$ and hypothesized effect size. This is called an *a priori* power analysis. The most difficult part is the specification of the population effect size. For a particular test, the effect size indicates the degree to which the null-hypothesis is false in the population. Since this population value is in general unknown, the effect size can be understood as the smallest departure from $H_0$ that we want to be able to detect with a given probability.

For a broad variety of statistical tests, Cohen (1988) presents indices of effect size, and procedures to determine the sample sizes needed to reach a specified power. Since researchers often have only a vague idea of what constitutes a plausible effect size, Cohen also proposes conventions that define 'small', 'medium', and 'large' effect sizes. For instance, for testing a correlation, the effect size is simply the value of the correlation in the population, and Cohen (1988, 1992) proposes .10, .30, and .50 as conventions for a 'small', 'medium', and 'large' correlation. These values correspond to 1%, 9%, and 25% of explained variance. Cohen remarks that a small effect is of a size that needs statistical analysis to detect it, while a medium effect is an effect size that one would become aware of given daily experience. A large effect is an effect that is immediately obvious.

The general procedure for estimating the power of a statistical test is illustrated in Figure 12.1. Let us assume that we have a test that results in a standardized $Z$-test

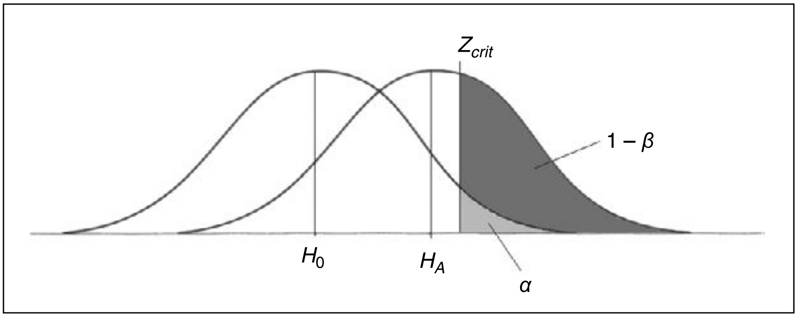


*Figure 12.1* Significance and power in the $Z$-test.

statistic. Under the null-hypothesis, $H_0$, the critical value for a one-sided test at $\alpha = .05$ is $Z_{crit} = 1.65$. Under the alternative hypothesis, $H_A$, we have a $Z$ distribution, which also has a variance equal to 1, but its mean is shifted by $\delta$ = (effect size)/(standard error). This distribution is known as the noncentral $Z$ distribution, with noncentrality parameter $\delta$, which in this case is simply the mean of the $Z$ distribution under $H_A$. The power of the test is the probability of exceeding the critical value $Z_{crit} = 1.65$ in the noncentral $Z$ distribution.

For example, the standard error of a correlation is approximately $1/\sqrt{N-3}$. If we assume a medium effect, the correlation in the population is assumed to be .30. For a sample of 50, the standard error of the correlation is approximately .15. The power of our test is the probability of exceeding $Z_{crit} = 1.65$ in the $Z$ distribution with mean .30/ .15 = 2. This is equal to the probability of exceeding $Z = 1.65 - 2 = -0.35$ in the standard normal $Z$ distribution, which turns out to be 0.64. Thus, our test has a power of .64.

A convenient formula for power analysis is (Snijders & Bosker, 1999, p. 142):

$$\text{(effect size)/(standard error)} \approx (Z_{1-\alpha} + Z_{1-\beta}) \tag{12.3}$$

Equation 12.3 contains four quantities. If three of them are known, we can compute the fourth one. So, in our example, if we want a power of .80, which means that $\beta = 0.20$, to detect a correlation of .30 the equation is:

$$\text{(.30)/(standard error)} = (Z_{0.95} + Z_{0.80}) = (1.64 + 0.84) = 2.48.$$

Thus, we need a standard error of (.3/2.48 =) .12. Calculating back to the sample size $N$ from the formula $s.e._r = 1/\sqrt{N-3}$, we find we need at least $((1/0.12)^2 + 3 = 8.33^2 + 3 =)$ 72.4, or a rounded 73 cases to achieve a power of .80.

To use equation 12.3 for power analysis, we need to know the standard error. This is easily found if we have analyzed available data and wish to assess the power of a specific test. This is *post hoc* power analysis. Since this is a simple case, it will be treated first with an example in section 12.2.1, and the more difficult case of a priori power analysis is treated in section 12.2.2.

### 12.2.1 Post hoc power analysis

Chapter 11 contains the results of a meta-analysis using multilevel analysis techniques. Table 11.4, which is repeated here as Table 12.0, presents the results of separately adding three explanatory study characteristics to the model. One of these is each study's total sample size $N_{tot}$. This variable is important, because if it shows a relationship with the effect size of the studies, this indicates that there may have been selective

*Table 12.0* Table 11.4, Multilevel meta-analyses on example data (repeated)

| Model | Intercept-only | + $N_{tot}$ | + reliability | + duration | + all |
|---|---|---|---|---|---|
| Intercept | 0.58 (.11) | 0.58 (.11) | 0.58 (.11) | 0.57 (.08) | 0.58 (.08) |
| $N_{tot}$ | | 0.001 (.01) | | | –0.00 (.01) |
| Reliability | | | .51 (1.40) | | –.55 (1.20) |
| Duration | | | | 0.14 (.04) | 0.15 (.04) |
| $\sigma_u^2$ | 0.14 | 0.16 | 0.16 | 0.04 | 0.05 |
| $p$-value $\chi^2$ deviance | $p < .001$ | $p < .001$ | $p < .001$ | $p = .15$ | $p = .27$ |
| $p$-value $\chi^2$ residuals | $p < .001$ | $p < .001$ | $p < .001$ | $p = .09$ | $p =.06$ |

publication. The relationship is in fact not significant, with a regression coefficient estimated as 0.001 and a standard error of 0.01. However, there is the problem of power. There are only 20 studies, so it is possible that this nonsignificance is the result of low power, rather than the absence of a real effect in the population.

To investigate this possibility, we must specify what effect size we wish to be able to detect. For a correlation coefficient, a medium effect is defined as .30, or 9% explained variance. For a regression model, we may consider a predictor to have a medium effect if it explains 10% of the variance, which is a nice round figure. In the intercept-only model in Table 12.0, the between-studies variance is estimated as $\sigma_u^2$ = 0.14. Ten percent of that variance equals 0.014, and that variance must be explained using a term of the form $\gamma N_{tot}$. In our data set in Table 12.0, we can calculate the variance of the total sample size $N_{tot}$, which turns out to have a value of 155.305. To reduce that to 0.014, the regression coefficient gamma must be equal to $\gamma = \sqrt{0.014}/\sqrt{155.305} = 0.01$. Therefore, we want to test an effect size of $\gamma = .01$, with an associated standard error of .01 (the value of the standard error for $N_{tot}$ in Table 12.0), and the significance level set at $\alpha = .05$. We again use equation 12.3: (effect size)/(standard error) ≈ $(Z_{1-\alpha} + Z_{1-\beta})$, which in this case becomes $(.01)/(.01) = (1.64 + Z_{1-\beta})$. So, $Z_{1-\beta} = 1 - 1.64 = -0.64$. This leads to a post hoc power estimate of .74, which appears adequate. The failure to find a significant effect for the study sample size is not likely to be the result of insufficient power of the statistical test.

Post hoc power analysis is not only useful in evaluating one's own analysis, as just shown, but also in the planning stages of a new study. By investigating the power of earlier studies, we find what effect sizes and intraclass correlations we may expect, which should help us to design our own study.

### 12.2.2 A priori power analysis: General issues

In an a priori power analysis, we want to estimate the power of our test for a specific effect size. Typically, we want to assess what sample size we need to achieve, say, a power of .80. In multilevel regression analysis, several factors complicate things. First, the multilevel regression model is a complex multivariate model, and to specify effect sizes and calculate standard errors we have to consider plausible values for many parameters of that model. Second, we have sample sizes at different levels. The same or similar power values may be obtained with different combinations of numbers of groups and group sizes. To decide which of these are going to be used, we must consider the cost of collecting data from more group members, or of collecting one more group. For example, assume that we want to assess the effect of an anti-smoking program, which is offered in school classes. In many cases, school research uses written questionnaires for data collection. In this case, once a class is selected for the study, it makes sense to collect data from all pupils in that class, since the extra cost of selecting one more pupil in a class is very low. Therefore, if the average class size is 20, we may decide to collect data from 10 experimental and 10 control classes, which will give us 400 pupils. On the other hand, if the data collection is done by computer, and the teacher has to send the pupils one by one to the computer to respond to the questionnaire, the cost (if only in time) of collecting one more pupil in a selected class is considerable. It may be better to select only 10 pupils at random in each class, and compensate by collecting data from a total of 40 classes. In fact, since the intervention is done at the class level, which means that the variable that codes for the intervention has no within-class variance, collecting data from more classes would certainly increase the power of our test. Since there are always cost considerations, the question of the best sample size always involves decisions about the optimal design.

For a large variety of classical statistical tests, Cohen (1992) presents procedures and tables that can be used for an a priori power analysis. For some special cases, similar procedures can be derived for power estimation in multilevel models. For instance, Cohen (1998) treats the problem of analyzing power in cluster samples for the case where we have only lower-level predictors and no random slopes. However, this case is more easily approached using the standard tables in Cohen (1992) and correcting the results for the 'design effect' (Kish, 1965, 1987). This correction procedure, described in Chapter 1 of this book, uses the intraclass correlation between respondents within clusters to correct the sample size:

$$n_{eff} = n/[1 + (n_{clus} - 1)\,\rho] \tag{12.4}$$

In equation 12.4, $n_{eff}$ is the effective sample size, $n_{clus}$ is the cluster size, $n$ is the total sample size, and $\rho$ is the intraclass correlation. So, if we know the intraclass correlation, or can make a reasonable guess about its size, we can find the cluster size and the total sample size needed to give us a specific effective sample size. By rearranging the correction formula, we get:

$$n = n_{eff}(1 + (n_{clus} - 1)\,\rho) \tag{12.5}$$

Earlier, we calculated the sample size needed to detect a correlation of .30 with a power of .80 as 73 cases. Suppose that we collect our data using three school classes. We have no information about the intraclass correlation $\rho$, but school research often reports an intraclass correlation $\rho = .1$. Taking that as a reasonable approximation, and using (73/3 =) 24.3 as the average class size, we find that with clustered data we need 243.3 or minimally 244 pupils, which implies about 10 classes. This is considerably more than the 73 pupils who are needed if the data are collected as a simple random sample! Summarizing: When we have only a random intercept and no random slopes, we can use Cohen's (1992) procedures to estimate the power, and follow this by adjusting the sample size for the design effect. For group-level explanatory variables, the intraclass correlation is equal to 1.0, and the effective sample size is equal to the number of groups. Again, we can use Cohen's (1992) tables to assess the power.

Raudenbush and Liu (2000) treat the important problem of determining the optimal design for multisite randomized trials, where experimenters combine the results from an experiment conducted at a number of different sites. This is an important design in biomedical research, where the number of patients in any specific location may be too small to be useful. By combining data from experiments carried out at a number of different sites, the power of the statistical analysis of the treatment can be increased. Since the patients are nested within the sites, multisite studies have a multilevel structure (Woodruff, 1997). Raudenbush and Liu (2000) consider the problem of determining the power to confirm a treatment effect, given the possibility that the treatment effect varies across sites. In multilevel terminology, they assume that the treatment variable has a random slope across sites, and include potential cross-level interactions to model the influence of site characteristics on the treatment effect. Raudenbush (1997) treats the same problem when the intervention is not implemented within the sites, but at the site or cluster level. This means that some sites are treatment sites, and others are control sites. In this design, the intraclass correlation of the treatment variable is 1.0. Both approaches to power estimation are not completely general, because they do not include individual-level explanatory variables. Snijders and Bosker (1993) present asymptotic formulas for the standard errors of fixed effects in two-level designs. The formulas are complicated, and they recommend using their program PinT (*P*ower *in* *T*wo-level models; Bosker, Snijders, & Guldemond, 2003) for the calculations. The power problems considered by Raudenbush (1997) and

Raudenbush and Liu (2000) can both be analyzed using PinT, but PinT also allows power calculations for the variance estimates.

When we consider power problems in multilevel designs, it is useful to translate the model under consideration into a standardized model. In a standardized model, the logic of power analysis, including decisions about effect sizes, does not depend on the arbitrary scales of the variables. In a standardized framework, it also becomes possible to establish rules of thumb about what may be considered small, medium, and large effects. Raudenbush and Liu (2000) propose standardizing the lowest-level variance to $\sigma_e^2 = 1.0$. If there is an experimental treatment, we can imagine that the treatment variable is coded using a dummy variable with values –0.5 and +0.5. This means that the intercept and the variance components reflect the average subject, and that the regression coefficient for the treatment dummy variable is equal to the standardized effect size *d* discussed in the meta-analysis chapter (Chapter 11) in this book. In a two-group experiment, the effect size *d* is defined as $d = (\bar{Y}_E - \bar{Y}_C)/s$, which is precisely what we get if the within-groups variance is standardized. Cohen (1988) defines $d = .2$ as a small effect, $d = .5$ as a medium effect, and $d = .8$ as a large effect. These values roughly correspond to the correlation values of .10, .30, and .50 mentioned earlier, and they can be used to find matching values for the regression coefficients if we have continuous variables. If we define a small effect as an effect that explains 0.01 of the variance, for a predictor variable with variance 1.0 this implies a regression coefficient of 0.10. This adds $0.1^2 = 0.01$ to the total outcome variance, which leads to $0.01/(1.00 + 0.01) = 0.01$ explained variance. In general, the standardized regression coefficient we seek is $b = \sqrt{r^2/(1 - r^2)}$, where *b* is the regression coefficient, and *r* is the corresponding correlation. Similarly, a correlation of .30 between the predictor variable and the outcome variable, with lowest-level error variance fixed at 1.0, implies a regression coefficient of 0.31, and a correlation of .50 between the predictor variable and the outcome variable, with lowest-level error variance fixed at 1.0, implies a regression coefficient of 0.58. As a result, when we want to evaluate the power of a test of a fixed regression coefficient using standardized continuous explanatory variables, and $s_e^2 = 1.0$, conventional criteria lead to values of .10 for a small, .31 for a medium, and .58 for a large effect size.

In multilevel regression, we must also be concerned with the intraclass correlation and the variance of the random slopes. In multilevel regression analysis, it is useful to have rules of thumb about the variance components. Raudenbush and Liu (2000) suggest values of 0.05, 0.10, and 0.15 as small, medium, and large variances for the slope of an intervention dummy coded –0.5, +0.5. Such suggestions are of course tentative and study-specific, but these values seem reasonable. For instance, Cohen (1988, 1992) views an intervention effect of 0.5 as a medium effect size. If the corresponding regression coefficient $\gamma = 0.5$ has a slope variance of 0.05, this translates to a standard deviation of 0.22. Assuming normality, 95% of the slopes would be between

0.06 and 0.94.[1] The combination of a medium treatment effect and a small variance across treatment sites leads to the result that virtually all population intervention effects are positive. A small intervention effect of 0.20 combined with a medium slope variance of 0.10 leads to an interval for the intervention effects across sites of between –0.42 and 0.82. In this situation, 26% of the intervention outcomes are expected to be zero or negative. This clearly underscores the importance of evaluating the treatment effect across sites. It also supports the rules of thumb chosen for small, medium, and large slope variances. The variance of the treatments across sites only becomes important if it has an effect size that is considerably larger than the effect size of the treatment.

For the variance of the intercept, the values of 0.05, 0.10, and 0.15 may be too small. They translate to intraclass correlations of .05, .09, and .13. In cluster sampling, variances are typically low (see Groves, 1989), and a value of 0.10 would be considered large in that context. In educational and organizational contexts, intraclass correlations of .10 seem reasonable, and .15 could indeed be considered high. However, in small group and family research, intraclass correlations are often much higher, and .15 would more likely be a medium effect, with .30 a large effect size. I suggest using .05, .10, and .15 as small, medium, and large values for the intraclass correlation in general cases, and .10, .20, and .30 in those cases where on a priori grounds much higher intraclass correlations appear reasonable. These translate to intercept variances of 0.05, 0.11, and 0.18 in ordinary situations, and 0.11, 0.25, and 0.43 in situations where a high intraclass correlation can be expected.

With these considerations and suggestions set for rules of thumb, the next sections explain a priori power analysis by examples.

### 12.2.3 A priori power analysis: The intercept variance and intraclass correlation

In applications such as meta-analysis, it is important to be able to detect between-study heterogeneity, or between-study variance. In the multilevel approach to meta-analysis (see Chapter 11 in this book), this translates to the significance of the second-level intercept variance. Longford (1993, p. 58) shows that the sampling variance of the second-level intercept variance $\sigma_u^2$ is equal to:

$$\mathrm{var}(\sigma_u^2) = \frac{2\sigma_e^4}{kn_{clus}}\left(\frac{1}{n_{clus}-1} + 2\omega + n_{clus}\omega^2\right) \tag{12.6}$$

[1] Note that this is not the 95% confidence interval. The 95% confidence interval relates to the precision of the estimate of the average effect across sites. In a random coefficient model, we assume variation around this average. The associated interval is the 95% predictive interval.

where $k$ is the number of groups or clusters, $n_{clus}$ is the cluster size, and $\omega$ is the ratio of the between- and within-studies variance: $\omega = \sigma_u^2 / \sigma_e^2$. Equation 12.6 can be used to estimate the power of detecting any specific between-study variance.

If we analyze standardized effect sizes, the first-level variance is implicitly fixed at 1.0. Following Raudenbush and Liu (2000), we may use values of 0.05, 0.10, and 0.15 as small, medium, and large between-study variances for the standardized effect. We wish to be able to detect a medium variance of 0.10 with a power of .80.

Suppose we plan a meta-analysis on a specific topic. We have carried out a computerized literature search, and found 19 references on our topic. Three are available in the local library. We collect these three studies, and code their effect sizes and total sample sizes. The results are in Table 12.1.

*Table 12.1* Three studies for an intended meta-analysis

| Study | $d$ | $N_{tot}$ |
|---|---|---|
| 1 | .03 | 256 |
| 2 | .12 | 185 |
| 3 | −.14 | 144 |

The question is: Given the results coded in the first three studies (actually, the first three studies in the meta-analysis example of Bryk & Raudenbush, 1992), is it worthwhile to go on, meaning retrieving and coding the remaining 16 studies? We can use an a priori power analysis to formulate an answer to this question. Specifically, assume that we wish to test whether the studies are heterogeneous, that is, whether the between-studies variance $\sigma_u^2$ is significant. We require a power of .80 to detect the study-level variance at the conventional $\alpha = .05$ significance level when the proportion of between-study variance is at least 0.25 of the total variance, a value that Hunter and Schmidt (2004) consider an important lower limit. Because we meta-analyze standardized effect sizes, the within-study variance is fixed at $\sigma_e^2 = 1.0$. For the proportion of between-study variance to be 0.25, the value of this variance must be $\sigma_u^2 = 0.33$ ($0.25 = 0.33/1.33$), and therefore $\omega = 0.33$. With $k = 19$ studies, an average study sample size $n_{clus} = 195$ (the average in Table 12.1), and $\omega = 0.33$, we obtain $\text{var}(\sigma_u^2) = 0.012$. Thus, the standard error of the second-level variance estimate $\sigma_u^2$ is 0.11. Using formula 12.3, we find that the power of the test of $\sigma_u^2 = 0.33$, where the standard error is .11, is estimated as (assuming a one-sided test: $p(Z > (1.64 - 0.33/.11) = p(Z > 0 - 1.31))$ .91, which appears more than adequate. If the sample sizes of the three available studies are typical for all studies, it appears worthwhile to continue the meta-analysis.

Similar calculations using the *design effect* (see section 12.2.2) allow us to assess the power of the overall test for the effect size *d*. Suppose we are interested in detecting a combined small effect, meaning an effect size as low as $\delta = .20$. This is small, but one of the advantages of meta-analysis is the possibility to detect small effects. If the sample sizes of the three available studies are typical for all studies, all studies together involve (19 × 195 =) 3705 subjects. However, we do not have one giant experiment with 3705 independent observations, we have 19 smaller experiments. We again assume that the between-studies variance is 0.25 of the total variance. In other words, we assume clustered data with 19 clusters of 195 subjects, and an intraclass correlation of .25. Using formula 12.4 to estimate the effective sample size from the intraclass correlation and the average cluster size, we obtain $n_{eff} = 3705/(1 + 194 \times .25) = 75$. Using the standard formula for the sampling error of the effect size (see Table 11.1 in Chapter 11), using 75 subjects with equal sample sizes for the experimental and control groups, we obtain an expected standard error for *d* of 0.23. Thus, the power estimate is (assuming a two-sided test: $p(Z > 1.96 - 0.10/0.077) = p(Z > 1.53)$) .06. We conclude that the power of our meta-analysis for detecting a small experimental effect is very poor. If we are interested in medium size effects, the power estimate is (assuming a two-sided test: $p(Z > 1.96 - 0.30/0.23) = p(\mathrm{Z} > 0.66)$) .25, which is again not adequate.

### 12.2.4 A priori power analysis: Designing an intervention study

Suppose we plan an intervention study to evaluate a course given to general practitioners (GPs) to improve their communication skills. Fifty GPs will participate in the study. In such studies, there is an issue of whether the randomization should take place at the patient level or at the GP level. In medical and organizational intervention studies, there is a similar issue of whether the randomization should be within or between organizational centers. In medical research, this issue is discussed under the heading of ‘cluster randomization’. Technically, randomization can be carried out at any of the available levels. Randomization within centers is generally more efficient, but it can also lead to ‘treatment-group contamination’, where information leaks from the experimental group to the control group (Moerbeek, 2000, p. 38). On the other hand, if randomization is at the organizational level, it is impossible to estimate a random component for the treatment variable. In our case, we choose a design in which all GPs are evaluated by a number of their patients, then they all follow a communication skill course, and after that they are evaluated again, using the same number of patients as before. Assuming the patients arrive at random, we have randomization at the patient level. The question is: How many patients do we need to interview?

To answer this question, we must first decide on the effect size, and the desired power. For this example, we decide that a medium effect is the largest that we may reasonably expect from our intervention. We decide on a minimal power of .80 for a medium effect, and .60 for a small effect. The significance level is $\alpha$ = .05 and to maximize the power we will carry out a one-tailed test. For the moment, we assume that the second-level sample size of 50 GPs is fixed by the study design, and we can increase the power only by using larger samples of patients within GPs. The calculations all assume equal group sizes. Since this is a planned study, we may assume that the investigators will aim to obtain equal numbers of patients per doctor.

*Power of a simple* t-*test*

The simplest analysis for this design is a $t$-test for independent groups after the intervention, which totally ignores the cluster effect of the GPs. To assess the power of a $t$-test for the difference between two means, we use the standardized difference, $d$, given by $d = (\bar{x}_1 - \bar{x}_2)/s$. Table 12.2 presents the power of the ordinary $t$-test for different combinations of effect size and sample size for each period, assuming 50 GPs and a spread of 9–15 patients for each GP in each period.

The power values in Table 12.2 assume that the patients are independent across GPs. However, it is reasonable to expect a dependency between observations collected from the same GP. The result is of course estimates for standard errors that are too small, and spurious 'significant' results. The correction described by Kish (1965, p. 259) referred to in equation 12.4 computes the effective sample size in two-stage cluster sampling as $n_{eff} = n/[1 + (n_{clus} - 1)\rho]$, where $n_{eff}$ is the effective sample size, $n_{clus}$ is the cluster size, $n$ is the total sample size, and $\rho$ is the intraclass correlation.

*Table 12.2* Power of simple $t$-test, intervention study

| $n_{per}$ | $n_{GP}$ | Two-tailed | | One-tailed | |
|---|---|---|---|---|---|
| | | small, $d = .2$ | medium, $d = .5$ | small, $d = .2$ | medium, $d = .5$ |
| 450 | 9 | .86 | 1.00 | .91 | 1.00 |
| 500 | 10 | .88 | 1.00 | .94 | 1.00 |
| 550 | 11 | .91 | 1.00 | .95 | 1.00 |
| 600 | 12 | .93 | 1.00 | .97 | 1.00 |
| 650 | 13 | .95 | 1.00 | .97 | 1.00 |
| 700 | 14 | .96 | 1.00 | .98 | 1.00 |
| 750 | 15 | .97 | 1.00 | .99 | 1.00 |

At this point, it is simple to calculate the effective sample size for different situations. For instance, suppose that we take a sample of 50 doctors, each with 9 patients for each period. This comes to a total sample size of 450 for each period, called $n_{per}$ in Table 12.2, which is reasonable. Let us further suppose that we are interested in a variable for which the intraclass correlation $\rho$ is .10. We have defined this earlier, following Raudenbush and Liu (2000), as a medium size intraclass correlation. It is a reasonable value, which has also been found in other research on patients clustered within GPs. However, using Kish's formula, the effective sample size in this situation is 250, which is much less than the apparent sample size of 450!

Table 12.3 presents the power of the ordinary $t$-test assuming an intraclass correlation of $\rho = .10$. It makes clear that in the presence of an intraclass correlation of .10, the effective sample size, given a fixed sample of 50 GPs, is relatively low, which leads to relatively low power values. However, medium effects can still be detected with ease in all cases, and even small effects have a reasonable probability of being detected. It appears that using only nine or ten patients for each GP is sufficient to meet our stated power criteria of .80 for a large and .60 for a small intervention effect. Using a one-sided test for the intervention effect is advisable if detection of small effects is important.

*Table 12.3* Power of $t$-test, cluster sample, intervention study

| $n_{per}$ | $n_{GP}$ | $n_{eff}$ | Two-tailed | | One-tailed | |
|---|---|---|---|---|---|---|
| | | | small, $d = .2$ | medium, $d = .5$ | small, $d = .2$ | medium, $d = .5$ |
| 450 | 9 | 250 | .61 | 1.00 | .72 | 1.00 |
| 500 | 10 | 263 | .63 | 1.00 | .74 | 1.00 |
| 550 | 11 | 275 | .65 | 1.00 | .76 | 1.00 |
| 600 | 12 | 286 | .67 | 1.00 | .77 | 1.00 |
| 650 | 13 | 295 | .68 | 1.00 | .78 | 1.00 |
| 700 | 14 | 304 | .69 | 1.00 | .79 | 1.00 |
| 750 | 15 | 312 | .70 | 1.00 | .80 | 1.00 |

### *Power in a variance component model*

A test for the effectiveness of the intervention can conveniently be carried out using multilevel analysis, with patients defined as level 1, and GPs as level 2. The simplest multilevel model is a variance component model, which assumes a fixed effect size for the intervention for all GPs, but allows for differences between GPs in general effectiveness. If only a dummy variable for the intervention is included, this analysis is equivalent to the *t*-test with correction for the design effect. Since the multilevel analysis does not assume equal group sizes, it is slightly more accurate if group sizes are different. Otherwise, the power should be similar to the estimates in Table 12.3.

If we assume that the treatment effect is the same for all GPs, but allow for an intraclass correlation, we have a variance component model. Multilevel analysis of this model has the advantage that it allows simple modeling of additional covariates at the patient or GP level, similar to adding covariates in a multiple regression model. In both periods, the same GPs participate. If in both periods the same kinds of patients visit their GPs, we have effective randomization of patients over treatment and control period. This implies a correlation between patient characteristics and the intervention dummy that is close to zero. As a result, inclusion of a patient variable that explains significant residual variance leads to a larger power for the test of the intervention dummy, because the residual variance $s_e^2$ becomes smaller (Cohen, 1988). Furthermore, if GP-level variables are added that explain variance, the partial intraclass correlation decreases, which (see the Kish equation) leads to a larger $n_{eff}$, and hence to a higher power. The increase in power can be estimated using procedures described by Cohen (1988), by including a correction for the design effect. These power estimates depend of course on ad hoc assumptions about the amount of variance explained and the correlations between the predictors. However, under realistic assumptions, adding effective explanatory variables at either the patient or the GP level leads to a higher power. Since the procedures described by Snijders and Bosker (1993) and later in this chapter include the variance component model as a special case, we will not look into the variance component model in more detail here.

### *Power in a random coefficient model*

It is realistic to assume that the effect of the intervention may differ across GPs, which leads to a multilevel model with a random slope coefficient for the intervention variable. The power of the test for the significance of the intervention coefficient in such models is addressed by Raudenbush and Liu (2000). They consider a multisite trial, where the average impact of an intervention is assessed across sites, and the moderating effect of site variables on the treatment efficacy is also assessed. As mentioned before, Raudenbush and Liu propose 0.05 for a small, 0.1 for a medium, and 0.15 for a large slope variance. They note that when the intervention effect varies a lot across sites, the average effect of an intervention is a poor indicator of the importance of the treatment effect, and masks possible negative effects of the intervention.

Raudenbush and Liu note that the average treatment effect can be tested using an *F*-test. Under the null-hypothesis *F* has an *F* distribution described by $F(1, J - 1)$, with the degrees of freedom for the denominator $J - 1$ equal to the number of groups minus one, and under the alternative hypothesis *F* follows a noncentral *F*-distribution *F*(1, *J* – 1; $\lambda$), with $\lambda$ the noncentrality parameter. This is analogous to the situation depicted in Figure 12.1, with the normal distribution replaced by the *F*-distribution. The noncentrality parameter is given by (following our notation):

$$\lambda = \frac{n_{tot}d^2}{n_{tot}\sigma_{u1}^2/J + 4\sigma_e^2} \tag{12.7}$$

where $\sigma_{u1}^2$ is the slope variance, and $\sigma_e^2$ is the individual-level residual variance. To find the power of the *F*-test, we must find the probability of exceeding the critical value for *F* in the noncentral *F*-distribution with noncentrality parameter $\lambda$ and degrees of freedom 1 and *J* – 1. Standard software packages like SPSS or SAS can be used to find this probability, or a separate program like the freely available NCSS Probability Calculator (NCSS, 1995).

Using Raudenbush and Liu's approach, we set up a power table for the intervention effect for two different postulated variances of the intervention slope across GPs. Table 12.4 again shows that for a medium effect all conditions result in sufficient power.

*Table 12.4* Power of test on intervention, random slope model

| $n_{tot}$ | Var slope | Two-tailed | | One-tailed | |
|---|---|---|---|---|---|
| | | small, *d* = .2 | medium, *d* = .5 | small, *d* = .2 | medium, *d* = .5 |
| 450 | 0.05 | .50 | 1.00 | .63 | 1.00 |
| 450 | 0.10 | .47 | 1.00 | .60 | 1.00 |
| 500 | 0.05 | .54 | 1.00 | .67 | 1.00 |
| 500 | 0.10 | .50 | 1.00 | .63 | 1.00 |
| 550 | 0.05 | .58 | 1.00 | .70 | 1.00 |
| 550 | 0.10 | .53 | 1.00 | .66 | 1.00 |
| 600 | 0.05 | .61 | 1.00 | .73 | 1.00 |
| 600 | 0.10 | .56 | 1.00 | .68 | 1.00 |
| 650 | 0.05 | .64 | 1.00 | .75 | 1.00 |
| 650 | 0.10 | .58 | 1.00 | .71 | 1.00 |
| 700 | 0.05 | .67 | 1.00 | .78 | 1.00 |
| 700 | 0.10 | .61 | 1.00 | .73 | 1.00 |
| 750 | 0.05 | .69 | 1.00 | .80 | 1.00 |
| 750 | 0.10 | .63 | 1.00 | .74 | 1.00 |

For detection of a small effect, most power estimates are unsatisfactory. Using a one-tailed test improves the power appreciably.

The structure of the equation that leads to Table 12.4 again leads to the conclusion that in order to increase the power, increasing the sample of GPs is more effective than increasing the sample of patients. Inclusion of a patient variable that explains significant residual variance leads to a larger power for the test of the intervention dummy. Furthermore, if GP-level variables are added that explain residual variance, the partial intraclass correlation decreases, which again (see the Raudenbush & Liu equation) leads to a higher power. The increase in power can be estimated, but these power estimates again depend to a large degree on ad hoc assumptions about the amount of variance explained and the correlations between the predictors. Under realistic assumptions, adding effective explanatory variables at either the patient or the GP level leads to a higher power. Including GP variables improves the power primarily if they explain intervention slope variance $\sigma_{u1}^2$.

We have already concluded that the power of our intervention is sufficient, even without covariates. However, it is an interesting question as to whether adding covariates would give us a minimal power of .80 even for small intervention effects. To assess the effect of explanatory variables at either the patient or the GP level, we need an approach that is more general than the one described by Raudenbush and Liu (2000). Snijders and Bosker (1993) describe a very general approach to estimating standard errors and power for fixed effects in two-level models, based on asymptotic approximations to the necessary standard errors. If we restrict ourselves to the fixed effects, the models described by Raudenbush and Liu are a special case of the general model considered by Snijders and Bosker. However, using it requires specifying and entering into the program PinT (Bosker et al., 1996) information about the means and the variances and covariances at both levels, for all predictor variables, and the variances and covariances of all random effects. As mentioned earlier, this requires a number of ad hoc assumptions about the amount of variance explained and the correlations between the predictors. Since patient and GP characteristics that are included as covariates correlate among themselves, only a limited number of covariates are useful. If there are numerous covariates, we can simplify the model by combining all covariates at each level in a single predictor variable. Thus, we simplify the model by assuming that we can find at least one (combined) patient-level and one (combined) GP-level variable that has at least a small effect in explaining the intercept variance at its own level.

Since we use a standardized model, we can specify all means as equal to zero. The within-patients variance $\sigma_e^2$ is standardized at 1.0. So far, we have assumed an intraclass correlation of .10, which translates to a second-level intercept variance of 0.11. PinT does not ask for the regression coefficients, but requires specification of the variances of the explanatory variables and of all random effects in the model.

Furthermore, explanatory variables at the lowest level are assumed to be group mean centered; their between-groups variance must be specified by adding the aggregated group means as a different variable.

We have a treatment effect that is coded –0.5/+0.5, which is assumed to have a small effect, meaning a regression coefficient $\gamma = .2$. The treatment dummy has a variance of 0.25, which is all within-GP variance.

We have a variance component model, with a fixed treatment effect and an intraclass correlation of .10. We can use a *t*-test, and adjust the sample sizes for the cluster effect. Table 12.5 presents power estimates for a one-sided test of a small effect ($\delta = .2$) at $\alpha = .05$, using the *t*-test adjusted for the design effect and using PinT. Although the power estimates are not exactly the same, they are close.

*Table 12.5* Comparison of different power estimation methods, variance component model

| $n_{tot}$ | $n_{GP}$ | Adjusted *t*-test | PinT estimate |
|---|---|---|---|
| 450 | 9 | .72 | .69 |
| 500 | 10 | .74 | .73 |
| 550 | 11 | .76 | .76 |
| 600 | 12 | .77 | .79 |
| 650 | 13 | .78 | .82 |
| 700 | 14 | .79 | .84 |
| 750 | 15 | .80 | .86 |

Using PinT, we can expand our model by allowing for a varying treatment effect across the GPs. This is equivalent to the problem addressed by Raudenbush and Liu (2000). Assuming a value of 0.1 for the slope variance (a medium effect size), and zero covariance between the intercept and the slope, we can calculate the power using either Raudenbush and Liu's approach or using PinT. The results are in Table 12.6. The two sets of power estimates are very close.

The next question is whether the inclusion of patient- or doctor-level explanatory variables improves the power. First, assume that it is possible to find an explanatory variable at the patient level that has a medium correlation with the outcome variable 'patient evaluation'. This means that 10% of the variance of the outcome variable is explained, and as a result, we expect the residual variance $\sigma_e^2$ to decrease to 0.90. The second-level variance $\sigma_{u0}^2$ will probably also decrease, let us assume also 10% to 0.10. Second, assume that we also find a GP-level variable that explains 10% of the

*Table 12.6* Comparison of different power estimation methods, random slope model

| $n_{tot}$ | $n_{GP}$ | Raudenbush & Liu | PinT estimate |
|---|---|---|---|
| 450 | 9 | .60 | .61 |
| 500 | 10 | .63 | .64 |
| 550 | 11 | .66 | .67 |
| 600 | 12 | .68 | .69 |
| 650 | 13 | .71 | .72 |
| 700 | 14 | .73 | .74 |
| 750 | 15 | .74 | .76 |

variance of the intercepts, so $\sigma^2_{u0}$ decreases further to 0.09. Experience shows that it is often difficult to explain slope variance, but let us assume that our GP-level variable also explains 5% of the slope variance, which then decreases from 0.10 to .095. To explain slope variance, we must build an interaction variable, so there are now four explanatory variables in the model: the intervention dummy, one patient variable, one GP variable, and the interaction term. To use PinT, we must specify all their (co)variances and the distribution over the two levels. The intervention variable has variance 0.25, at the patient level only. We assume that all other variables are standardized. The GP variable has all its variance at the GP level. If we assume an intraclass correlation of .10 for the outcome variable, we may as well assume the same for the explanatory patient variable, which gives us a patient-level variance of 0.90 and a GP-level variance of 0.10. The interaction variable is a multiplication of a variable with only patient-level variance (0.25) and a variable with only GP-level variance (1.0). We assume that, after standardization, its variance at the patient level is 0.20, and 0.80 at the GP level. Since we have approximate randomization, we assume no correlation between the intervention dummy and the other explanatory variables. For simplicity, we also assume zero correlations between the other variables. Only the interaction term correlates 0.60 with the two constituting variables. Since PinT expects covariances, we must calculate the ensuing covariances at both levels.

This is quite a bit of work, and it is clear that we must make some informed guesses and plausible simplifications regarding the explanatory variables. However, we do not need great precision here. The goal in a power analysis is to investigate whether adding explanatory variables is worthwhile for our objective of testing the intervention effect. Plausible values for these variables are sufficient for that goal. If we have different objectives, for instance if we are interested in the power for testing the interaction effect, we must be more specific about these values, and their justification.

Table 12.7 compares the results of the two PinT power estimates. It shows only modest increases in the power of the test of the intervention. Therefore, we conclude that adding covariates at the patient or GP level makes sense only if the added cost of the data collection is small.

*Table 12.7* Comparison of power, random slope model, without and with covariates

| $n_{tot}$ | $n_{GP}$ | Random slope, no covariates | Random slope plus covariates |
|---|---|---|---|
| 450 | 9 | .61 | .64 |
| 500 | 10 | .64 | .68 |
| 550 | 11 | .67 | .70 |
| 600 | 12 | .69 | .73 |
| 650 | 13 | .72 | .75 |
| 700 | 14 | .74 | .77 |
| 750 | 15 | .76 | .79 |

### 12.2.5 A general procedure for power analysis

All power estimation procedures discussed so far proceed by estimating the standard error of a specific parameter, and then applying equation 12.3, or rather its reverse:

$$Z_{1-\beta} = \frac{\text{effect size}}{\text{standard error}} - Z_{1-\alpha} \tag{12.8}$$

The various formulas assume that the data are balanced, meaning that the group sizes are equal, and that any randomization into experimental and control groups results in a 50:50 ratio. Deviations from these assumptions generally lead to designs with less power, but unless the deviations are large the formulas yield good approximations.

We may have a design that is not covered by the models discussed earlier, or we may have very unbalanced data. There is a very general approach for estimating the required standard errors, comparable to the approach used to estimate power in structural equation models (see Satorra & Saris, 1985, and Satorra, 1989). This can be summarized in three steps:

1. Specify a model that contains all variables and parameters of interest, including the parameter that is to be tested. All parameters are fixed at their (assumed) population values.

2. Generate data that reflect the population parameter values *exactly*.
3. Analyze the generated data using one of the available multilevel programs, and allow all parameters to be estimated.

The parameter estimates obtained in step 3 should be identical to the fixed values in step 1. The standard errors produced by the program are the standard errors needed for power estimation.

The difficult step is step 2: generate the data. In regression analysis, the explanatory variables are assumed fixed, so in theory we should know them. In practice the explanatory variables are usually also sampled. However, their precise distribution is not crucial. Unless they are dummy variables, like the intervention dummy used earlier, it makes sense to generate them as standardized variables with a multivariate normal distribution. It is more difficult to generate the residuals, since they must follow a specific multivariate normal distribution exactly. The solution is (Bollen & Stine, 1992) to generate a set of (centered) residuals **Y** with an arbitrary multivariate normal distribution characterized by covariance matrix **S**, and transform these to multivariate normal distributed residuals **Z** having the desired covariance matrix **Σ**, using the transformation:

$$\mathbf{Z} = \mathbf{Y}\mathbf{S}^{-0.5}\mathbf{S}^{0.5}. \tag{12.9}$$

A much simpler but approximate solution is to generate the residuals directly from **Σ**, skipping the transformation given by 12.9, and repeating the whole procedure three to five times to obtain an average standard error from the simulated distribution.

To use the exact solution for our example, we must set up an empty data set that has 50 GPs, and a specific number of patients for each GP. The equation for the model including the GP and patient variables is:

$$Y_{ij} = \gamma_{00} + \delta D_{ij} + \gamma_{01} P_{ij} + \gamma_{10} G_j + \gamma_{11} D_{ij} G_j + u_{0j} + u_{1j} P_{ij} + e_{ij}, \tag{12.10}$$

where $D$ is the intervention dummy coded –0.5, +0.5, $P$ is the patient variable, and $G$ is the GP variable.

Again, we assume that the explanatory variable $P$ at the patient level has a medium correlation with the outcome variable ‘patient evaluation’. This means that 10% of the variance of the outcome variable is explained, and as a result, we expect the residual variance $\sigma_e^2$ to decrease to 0.90. The second-level variance $\sigma_{u0}^2$ also decreases by 10% to 0.10. The GP-level variable $G$ explains another 10% of the variance of the intercepts, so $\sigma_{u0}^2$ decreases further to 0.09. We assume that our GP-level variable also explains 5% of the slope variance, which then decreases from 0.10 to .095.

First, we must set up our simulated population data. The explanatory variables $P$

and *G* are generated independently from a standard normal distribution. If we want *P* to have an intraclass correlation of .1, we must generate a group-centered normal variable with a variance of 0.9, and a group-level normal variable with a variance of 0.1, and add these. Since all this data generation is a random process, the mean and variance of the generated values for *P* and *G* will not be exactly zero and one, so they are standardized, and subsequently the product *DG* is calculated. Without loss of generality, we can set the intercept term $gg_{00}$ to zero. In a standardized model, $\sigma_e^2$ is set to 1.0. The regression coefficient $\delta$ for the intervention dummy *D* is set to .2, to reflect a small effect. The regression coefficients $\gamma_{01}$ and $\gamma_{10}$ are set to .31, to reflect a medium effect (see section 12.2.1). The residual intercept variance $\sigma_{u0}^2$ is set to 0.09, the residual slope variance $\sigma_{u1}^2$ is set to 0.095, and their covariance is set to zero. If we standardize *DG*, explaining 5% of the slope variance implies a value of .22 for $\gamma_{11}$. The lowest-level residual variance $\sigma_e^2$ is set to 0.9. The residuals $u_0$ and $u_1$ are generated independently from a standard normal distribution. Again, since this data generation is a random process, their mean will not be exactly zero, and their variances will not be exactly equal to the theoretical values. In addition, their correlation will not be exactly zero, so we must apply the transformation given by 12.6 to make the residuals $u_0$ and $u_1$ follow the desired distribution exactly.

Estimating power using the general procedure outlined in this section implies much work. For problems that can be formulated within the PinT framework, using PinT is less complicated. For more general problems, it is convenient to use an approximate solution that generates the needed residuals directly from $\boldsymbol{\Sigma}$ and skips the transformation given by 12.9. Repeating the whole procedure three to five times to obtain an average standard error from the simulated distribution produces estimated standard errors that are accurate enough for an a priori power analysis (Hox, 2001).

# 13

# Advanced Issues in Estimation and Testing

The usual method to estimate the parameters of the multilevel regression model is maximum likelihood estimation. This produces parameter estimates and asymptotic standard errors, which can be used to test the significance of specific parameters, or to set a confidence interval around a specific parameter. Chapter 3 mentions alternatives to this standard approach to estimation and testing. This chapter discusses several alternatives in more detail: the profile likelihood method, robust standard errors, bootstrapping, and Bayesian methods.

To provide more insight into the details and show the effects of different estimation methods, two example data sets will be used throughout this chapter. The first is a small data set, containing 16 independent measurements of the estrone level in a single blood sample from five women (the data are described in Appendix A). This data set is presented and discussed by Fears, Benichou, and Gail (1996) to illustrate the fallibility of the Wald statistic (based on the parameter estimate divided by the estimated standard error) for testing variance components in certain situations. In these example data, the Wald test fails to detect a variance component for two reasons: first, because the sample size is small (Fears et al., 1996), and second, because the likelihood for the subject-level variance is decidedly non-normal (Pawitan, 2000). In addition to this data set, which is known to be problematic, the pupil popularity data introduced in Chapter 2 are used. This is a large data set (2000 pupils in 100 classes), which has been generated following the assumptions of the multilevel model using maximum likelihood. Given the generous sample size, this well-behaved data set should produce accurate estimates and standard errors for all estimation methods.

The estrone data file is restructured to create a ‘long’ or ‘stacked’ data file that contains the 16 measurements nested within the five subjects. Following Fears et al. (1996) the (10-base) logarithm of the estrone level is used as dependent variable. Assuming independence of the 16 measurements on the same blood samples, one-way random effects analysis of variance can be used to assess whether the five women have significantly different average estrone levels. Since the data set is balanced, standard analysis of variance methods (Stevens, 2009; Tabachnick & Fidell, 2007) produce exact results. An analysis of variance on the estrone data yields the results shown in Table 13.1.

The *F*-ratio is highly significant, providing strong evidence that estrone levels vary between individuals. Using the variance components as estimated by the analysis

*Table 13.1* Analysis of variance on estrone data, random effects model

| Source | *df* | *SS* | *MS* | Var. comp. | *F*-ratio | *p* |
|---|---|---|---|---|---|---|
| Subjects | 4 | 1.133 | 0.283 | 0.0175 | 87.0 | < .0001 |
| Error | 75 | 0.244 | 0.003 | 0.0022 | | |
| Total | 79 | 1.377 | | | | |

of variance method produces an intraclass correlation of $\rho$ = .84, which indicates that most of the variation in this data is between-subject variation. Multilevel analysis using maximum likelihood estimation with the restricted maximum likelihood (RML, SPSS) method should lead to similar estimates. The multilevel approach estimates the variance components as $s_{u0}^2 = 0.0175$ at the subject level and $s_e^2 = 0.00325$ at the measures (error) level. These estimates are close to the values obtained using analysis of variance, and the multilevel method produces an intraclass correlation of $\rho$ = .84. However, using the Wald test by dividing the variance estimate of 0.0175 by its estimated standard error of .0125 produces a standard normal variate $Z$ = 1.40, corresponding to a one-sided $p$-value of .081, which is not significant by conventional criteria. Clearly, the Wald test is not performing well with these data. The difference in the estimated variance components is trivial, so the problem is not the maximum likelihood estimation method, but the Wald test itself. The reason that the Wald test is performing badly in this example is simple. The Wald test depends on the assumption that the parameter tested has a normal sampling distribution, with a sampling variance that can be estimated from the information matrix. In the estrogen data, we are testing a variance component that does not have a normal distribution, under the conditions of a very small sample size, and close to its boundary value of zero.

Some simple alternatives discussed in Chapter 3 work well for these data. For instance, Longford (1993) and Snijders and Bosker (1999) suggest basing the Wald test not on the variance, but on the standard deviation $s_{u_0} = \sqrt{s_{u_0}^2}$, with standard error equal to $s.e.\left(s_{u_0}\right) = s.e.\left(s_{u_0}^2\right)/\left(2s_{u_0}\right)$. The standard deviation is a square root transformation of the variance, and its distribution should be closer to normality. For our data, $s_u$ is 0.132, with estimated standard error calculated as .0125/(2 × 0.132) = .047. A Wald test on the standard deviation produces a test value $Z$ = 2.81, with $p$ = .003. So this test indeed performs better. However, in the general case, solving the problem by applying some transformation to the estimated variance is problematic. Fears et al. (1996) show that, since the Wald test depends on the parameterization of the model, by making a shrewd choice of a power transformation for $s_{u0}^2$, one can obtain any $p$-value between 0 and 1. This is awkward, and better methods than transforming the variance estimate are preferable.[1]

If we use the chi-square test discussed in Chapter 3, and implemented in HLM, we find $\chi_4^2 = 348.04$, with $p < .001$. Similarly, if we use the deviance difference test discussed in Chapter 3, we find a deviance difference of 114.7 (RML). This is distributed as a chi-square variate with one degree of freedom, and can be converted to a standard normal $Z$-variate by taking its square root. This produces $Z = 10.7$, which is highly significant. In effect, both the residuals chi-square method and the chi-square test on the difference of the deviances work well on these data. However, these methods cannot be used if we wish to establish a confidence interval for the subject-level variance. The next sections discuss some alternatives to the Wald test that allow the construction of a valid confidence interval for the variance components.

## 13.1 THE PROFILE LIKELIHOOD METHOD

For the estrone data, the (RML, MLwiN) estimate for the intercept is 1.418 (s.e. = .06). The estimate for the subject-level variance $\sigma_{u0}^2$ is 0.0175. The deviance for the model is calculated as –209.86.[2] If we restrict this variance component to zero, the deviance becomes –97.95. It has gone up by a considerable amount, and the difference of 111.91 can be tested against the chi-square distribution with one degree of freedom. Using the deviance test, the variance component is clearly significant. Since the Wald procedure is suspect for these data, a 95% confidence interval for the subject-level variance based on the asymptotic standard error is also questionable. An alternative is a confidence interval that is based directly on the deviance test, similar to the procedures followed in the null-hypothesis test based on the deviance. Such a procedure exists, namely the *profile likelihood* method, and the resulting confidence interval is called a *profile likelihood* interval.

To construct a profile likelihood interval for the estrone data, we need a multilevel analysis program that allows putting constraints on the fixed and random parameters in the model. First, we constrain all parameters to their estimated values.

[1] It should be noted that HLM, MLwiN, and SPSS all produce slightly different results for the second-level variance and its standard error. When recalculated to a standard deviation and its associated standard error, these differences become exaggerated, which is another reason why this procedure is not recommended.

[2] This value is derived using RML in MLwiN; using RML in HLM produces a slightly different estimate. The difference is that MLwiN cannot calculate the RML deviance, so RML estimation is combined with values from the FML deviance. MLwiN is used throughout this chapter because it can impose constraints on variances, which is necessary here.

As a check, this should produce the same deviance as freely estimating them (within bounds of rounding error). Next, we constrain the value for the parameter that we wish to test to a different value. As a result, the deviance goes up. To reach significance, the increase in deviance must exceed the critical value in the chi-square distribution with one degree of freedom. For a 95% confidence interval, this critical value is 5.0239. So, to establish a 95% confidence interval around the subject-level variance estimate $s_{u0}^2 = 0.0175$, we must find an upper limit $\mathrm{U}\left(s_{u0}^2\right)$ for which the deviance is $-209.86 + 5.02 = -204.84$, and a lower limit $\mathrm{L}\left(s_{u0}^2\right)$ for which the deviance is also $-204.84$. These limits can be found by trial and error, or more efficiently by using a simple search method such as setting an interval that is on both sides of the limit we are looking for, and successively halving the interval until the limit is estimated with sufficient precision.

Using the profile likelihood method on the estrone data, we find a 95% confidence interval for $\sigma_{u0}^2$: $0.005 < \sigma_{u0}^2 < 0.069$. The profile likelihood confidence interval does not include zero, so the null-hypothesis of no subject-level variance is rejected. The profile likelihood interval is not symmetric around the estimated value of $\sigma_{u0}^2 = 0.018$. Of course, it is known that variance components follow a chi-square distribution, which is not symmetric, so a valid confidence interval for a variance component should also be non-symmetric.

## 13.2 ROBUST STANDARD ERRORS

When the response variable does not have a normal distribution, the parameter estimates produced by the maximum likelihood method are still consistent and asymptotically unbiased, meaning that they tend to get closer to the true population values as the sample size becomes larger (Eliason, 1993). However, the asymptotic standard errors are incorrect, and they cannot be trusted to produce accurate significance tests or confidence intervals (Goldstein, 1995, p. 60). This problem does *not* always vanish when the samples get larger.

Sometimes it is possible to obtain more nearly normal variables by transforming the outcome variable. If this is undesirable or even impossible, another method of obtaining better tests and intervals is to correct the asymptotic standard errors. One available correction method to produce robust standard errors is the so-called Huber/White or sandwich estimator (Huber, 1967; White, 1982). In maximum likelihood estimation, the usual estimator of the sampling variances and covariances is based on the Information matrix, or more general on the inverse of the so-called Hessian matrix (see Eliason, 1993). The standard errors used in the Wald test are simply the square root of the sampling variances that are found on the diagonal of this inverse. Thus, using matrix notation, the asymptotic variance–covariance matrix of the estimated regression coefficients can be written as:

$$\mathbf{V_A}\,(\hat{\beta}) = \mathbf{H}^{-1} \tag{13.1}$$

where $\mathbf{V_A}$ is the asymptotic covariance matrix of the regression coefficients, and $\mathbf{H}$ is the Hessian matrix. The Huber/White estimator is given as:

$$\mathbf{V_R}\,(\hat{\beta}) = \mathbf{H}^{-1}\mathbf{C}\mathbf{H}^{-1} \tag{13.2}$$

where $\mathbf{V_R}$ is the robust covariance matrix of the regression coefficients, and $\mathbf{C}$ is a correction matrix. In equation 13.2, the correction matrix is 'sandwiched' between the two $\mathbf{H}^{-1}$ terms, hence the name 'sandwich estimator' for the Huber/White standard errors. The correction term is based on the observed raw residuals. If the residuals follow a normal distribution, $\mathbf{V_A}$ and $\mathbf{V_R}$ are both consistent estimators of the covariances of the regression coefficients, but the model-based asymptotic covariance matrix $\mathbf{V_A}$ is more efficient since it leads to the smallest standard errors. However, when the residuals do not follow a normal distribution, the model-based asymptotic covariance matrix is not correct, while the observed residuals-based sandwich estimator $\mathbf{V_R}$ is still a consistent estimator of the covariances of the regression coefficients. This makes inference based on the robust standard errors less dependent on the assumption of normality, at the cost of sacrificing some statistical power. The precise form of the correction term is different in different models; for a technical discussion see Greene (1997). In multilevel analysis the correction is based on the cross-product matrix of the residuals, taking the multilevel structure of the data into account. Several multilevel packages can produce robust standard errors for the fixed part, and MLwiN and Mplus also use robust sandwich estimators for the standard errors of the variance components. The robust standard errors are related to the generalized estimating equations (GEE) estimation method described in Chapter 3, which uses the observed residuals to estimate the variance components in the model. In HLM, the robust standard errors automatically select GEE estimation.

When heteroscedasticity is involved because of non-normality, outliers, or misspecification of the model, the asymptotic standard errors are generally too small. Typically, the robust standard errors do not completely correct this, but they do result in more accurate significance tests and confidence intervals (Beck & Katz, 1997). So, when strong non-normality is suspected, it is prudent to use the sandwich standard errors. Since the robust standard errors are partly based on the observed residuals, they do need a reasonable level 2 sample size to be accurate; single-level simulation results by Long and Ervin (2000) suggest a sample size of at least 100. In multilevel analysis, this would translate to a minimum second-level sample size of 100 for the robust standard errors to work well. Multilevel simulations with strongly non-normal two-level data (Hox & Maas, 2001) confirm these recommendations. Cheong, Fotiu, and Raudenbush (2001) find that robust standard errors even provide reasonable protection against omitting an entire level from the analysis. On the other hand, robust standard errors tend to be larger than asymptotic standard errors (Kauermann & Carroll, 2001), so their routine use in situations where the assumptions are justified results in standard errors that are too large and hence in loss of power.

Since the sandwich estimator needs a reasonable sample size to work well, the estrone data with $N = 5$ are not a good example. We will use the pupil popularity data introduced in Chapter 2 to illustrate the use of sandwich standard errors. The model that we use is a random component model, with grand mean-centered predictors and FML estimation. By omitting the significant variance term for the slope of pupil extraversion, we introduce a misspecification in the model, which causes heteroscedasticity in the second-level residuals. Figure 13.1 shows a plot of the second-level residuals $u_0$ against their ranks in this model. There is indeed some evidence of non-normality at the extremes.

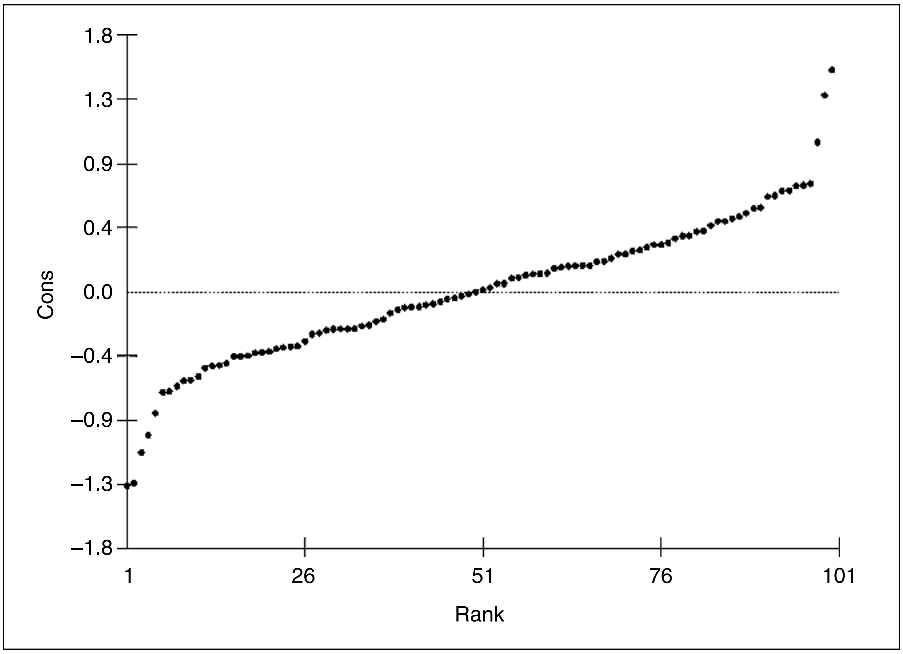


*Figure 13.1* Popularity data: Plot of second-level residuals against their rank.

Table 13.2 presents the parameter estimates, standard errors, and 95% confidence intervals using both the asymptotic and the sandwich standard errors. The parameter estimates are the same, and most of the standard errors and confidence

*Table 13.2* Comparison of asymptotic and robust results, popularity data

| | ML estimates, asymptotic s.e. (a.s.e.) | | ML estimates, robust s.e. (r.s.e.) | |
|---|---|---|---|---|
| **Fixed** | Coeff. (s.e.) | 95% CI | Coeff. (s.e.) | 95% CI |
| Intercept | 5.07 (.06) | 4.96–5.18 | 5.07 (.06) | 4.96–5.18 |
| Pup. gender | 1.25 (.04) | 1.18–1.33 | 1.25 (.04) | 1.18–1.32 |
| Pup. extrav. | 0.45 (.02) | 0.42–0.49 | 0.45 (.03) | 0.41–0.50 |
| Teacher exp. | 0.09 (.01) | 0.07–0.11 | 0.09 (.01) | 0.07–0.11 |
| **Random** | | | | |
| $\sigma_e^2$ | 0.59 (.02) | 0.55–0.63 | 0.59 (.02) | 0.55–0.63 |
| $\sigma_{u0}^2$ | 0.29 (.05) | 0.20–0.38 | 0.29 (.05) | 0.19–0.39 |

intervals are the same or very close. Only the robust standard error of the slope for pupil extraversion is larger. The 95% confidence interval, which has been calculated carrying more decimals, shows a small difference for pupil extraversion and for the class-level variance. Presumably, this reflects the misspecification caused by ignoring the random component for the extraversion slope.

An approach related to the sandwich estimators is the GEE (generalized estimating equations) approach of Liang and Zeger (1986). GEE estimation is a quasi-likelihood approach that starts by estimating the variance components directly from the raw residuals, followed by GLS estimation for the regression coefficients. This results in estimates for the regression coefficients that are consistent, but less efficient than maximum likelihood estimates (see Goldstein, 1995, p. 22; for a discussion of the GEE and other approaches see Pendergast, Gange, Newton, Lindstrom, Palta, & Fisher, 1996). If the second-level sample size is reasonable ($N > 100$, see Hox & Maas, 2001; Long & Ervin, 2000), the GEE estimates for the standard errors are not very sensitive to misspecification of the variance component structure. The software HLM uses GEE to estimate robust standard errors for the regression coefficients. Raudenbush and Bryk (2002, p. 278) suggest that comparing the asymptotic standard errors calculated by the maximum likelihood method to the robust standard errors is a way to appraise the possible effect of model misspecification and non-normality. Used in this way, robust standard errors become an indicator for possible misspecification of the model or its assumptions. If the robust standard errors are very different from the asymptotic standard errors, this should be interpreted as a warning sign that some distributional assumption has been violated, and as advice to look into the problem.

## 13.3 MULTILEVEL BOOTSTRAPPING

In its simplest form, the *bootstrap* (Efron, 1982; Efron & Tibshirani, 1993) is a method of estimating the parameters of a model and their standard errors strictly from the sample, without reference to a theoretical sampling distribution. The bootstrap directly follows the logic of statistical inference. Statistical inference assumes that in repeated sampling the statistics calculated in the sample will vary across samples. This sampling variation is modeled by a theoretical sampling distribution, for instance a normal distribution, and estimates of the expected value and the variability are taken from this distribution. In bootstrapping, we draw a sample *b* times (with replacement) from the observed sample at hand. In each sample, we calculate the statistic(s) of interest, and the observed distribution of the *b* statistics is used for the sampling distribution. Thus, estimates of the expected value and the variability of the statistics are taken from this empirical sampling distribution (Mooney & Duvall, 1993; Stine, 1989; Yung & Chan, 1999).

Since bootstrapping takes the observed data as the sole information about the population, it needs a reasonable original sample size. Good (1999, p. 107) suggests a minimum sample size of 50 when the underlying distribution is not symmetric. Yung and Chan (1999) review the evidence on the use of bootstrapping with small samples. They conclude that it is not possible to give a simple recommendation for the minimal sample size for the bootstrap method. However, in general the bootstrap appears to compare favorably over asymptotic methods. A large simulation study involving complex structural equation models (Nevitt & Hancock, 2001) suggests that, for accurate results despite large violations of normality assumptions, the bootstrap needs an observed sample of more than 150. Given such results, the bootstrap is not the best approach when the major problem is a small sample size. When the problem involves violations of assumptions, or establishing bias-corrected estimates and valid confidence intervals for variance components, the bootstrap appears to be a viable alternative to asymptotic estimation methods.

The number of bootstrap iterations *b* is typically large, with *b* between 1000 and 2000 (Booth & Sarkar, 1998; Carpenter & Bithell, 2000). If the interest is in establishing a very accurate confidence interval, we need an accurate estimate of percentiles close to 0 or 100, which requires an even larger number of iterations, such as $b > 5000$.

The bootstrap is not without its own assumptions. A key assumption of the bootstrap is that the *resampling* properties of the statistic resemble its *sampling* properties (Stine, 1989). As a result bootstrapping does not work well for statistics that depend on a very 'narrow feature of the original sampling process' (Stine, 1989, p. 286), such as the maximum value. Another key assumption is that the resampling scheme used in the bootstrap must reflect the actual sampling mechanism used to collect the data (Carpenter & Bithell, 2000). This assumption is of course very important in multilevel modeling, because in multilevel problems we have a hierarchical sampling mechanism, which must be simulated correctly in the bootstrapping procedure.

### 13.3.1 A simple example of bootstrapping

To demonstrate the basic principles of the bootstrap, it is convenient to use a small example, omitting the complications of multilevel data for the moment. Table 13.3 presents test scores of 12 students on an algebra test and a statistics test (Good, 1999, p. 89).

*Table 13.3* Algebra and statistics scores of 12 students

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Algebra | 80 | 71 | 67 | 72 | 58 | 65 | 63 | 65 | 68 | 60 | 60 | 59 |
| Statistics | 81 | 81 | 81 | 73 | 70 | 68 | 68 | 63 | 56 | 54 | 45 | 44 |

In this small data set, the correlation between algebra and statistics is $r = .677$. The two-sided $p$-value provided by the usual $t$-test, which is exact in small samples, gives a $p$-value of .016. We conclude that knowing algebra apparently helps in getting a good grade in statistics. If we want to calculate a 95% confidence interval for the correlation, we face two problems. First, it is known that the sampling distribution of correlations is not normal. Large correlations (such as .677) have a skewed distribution (Hedges & Olkin, 1985, p. 225). The usual solution for this problem is to use a Fisher-$Z$ transformation (see Hays, 1994).[3] Second, the standard error for the Fisher-$Z$, which is given by $1 / \sqrt{N - 3}$, is a large sample estimate. For our $N = 12$, the standard error is estimated as .333. If we apply the Fisher-$Z$ transformation, establish the 95% confidence interval for $Z$, and transform back to correlations, we find a 95% confidence interval for the correlation $r$ that runs from 0.151 to 0.897. The $p$-value based on the asymptotic method is $p = .042$. However, $N = 12$ is not a large sample, and the application of the large-sample standard error is questionable.

In a bootstrap approach, we resample the data set $k = 1000$ times, each time computing the correlation between algebra and statistics (using Amos, see Arbuckle, 2007). This produces 1000 correlations, with a mean of .668, and a standard deviation of 0.158. Establishing a confidence interval using the 1000 bootstrapped correlations is possible by means of two different techniques. One approach is to take the mean value

[3] Fisher's $Z$ is $Z = 0.5\ln((1 + r)(1 - r))$, and the inverse is $r = (e^{2Z} - 1)/(e^{2Z} + 1)$.

of the 1000 correlations as the point estimate, followed by establishing lower and upper limits for the 95% confidence interval using a normal approximation, with the standard deviation of the bootstrapped correlations as an estimate of the standard error (Mooney & Duval, 1993). When the statistic being bootstrapped is unbiased and follows an unbounded symmetric distribution, this method produces a valid confidence interval (Carpenter & Bithell, 2000; Stine, 1989). In our small example, the normal approximation produces a 95% confidence interval from 0.496 to 1.115. The normal approximation does not recognize boundaries such as 1.0 to the correlation; as a result, it produces an impossible value for the upper limit.[4] Since the sampling distribution of the correlation is skewed, the normal approximation is not a good choice. The other approach, which is much closer to the nonparametric spirit of the bootstrap method, is to use the 2.5th and 97.5th percentiles of the observed bootstrap distribution as limits of the 95% confidence interval (Mooney & Duval, 1993; Stine, 1989). Since this needs a precise estimate of the 2.5th and 97.5th percentiles, we should use a large number of bootstrap samples here, such as 4999. The 95% confidence interval for our example, using the percentile method, runs from 0.295 to 0.906, with a $p$-value of .009.

The application of the bootstrap method to obtain standard errors for parameter estimates and establishing confidence intervals is straightforward. If we could sample, say, 1000 real samples from our population, we could calculate the sampling variance directly. Since this is not possible, we use the computer to *resample* 1000 samples from our sample data. This simulates the actual sampling, which is in practice not possible, and provides a simulated estimate of the sampling variance. In addition to providing parameter estimates and sampling variances, there are some less obvious refinements to the bootstrap method. For instance, it is possible to use the bootstrap method to correct the asymptotic parameter estimates. The mean of the bootstrapped parameters is not necessarily equal to the estimate in the original sample. On the contrary, it can be rather different. If that is the case, the assumption is that the statistic under consideration is biased. Whatever mechanism is operating to produce bias in the bootstrap samples is assumed to be operating in the original sample as well. To correct for this bias, we use the difference between the original estimate and the mean bootstrap estimate as an estimate of the amount of bias in the original estimate. In our example, the correlation in the original sample is .677, and the bootstrap mean is .668. Thus, the original correlation has a negative bias that is estimated as –0.009. Using this bias estimate, the original correlation is corrected to .677 + 0.009 = .686. Although this is a minute correction, the ordinary correlation coefficient has indeed a small negative bias in small samples (see Hedges & Olkin, 1985, p. 225). With our very small sample, the bias becomes noticeable, and the bootstrap procedure can be used to

[4] Bootstrapping Fisher-$Z$ transformed correlations would solve this problem.

correct for it. Using a more complicated procedure (for details see Stine, 1989, and Mooney & Duval, 1993) the limits of the confidence interval can also be corrected for this bias. The 95% confidence interval using the bias-corrected percentile method runs from 0.234 to 0.892, with a *p*-value of .015. The bias-corrected or *BC p*-value is very close to the exact *p*-value of .016 given by the small-sample *t*-test.

### 13.3.2 Bootstrapping multilevel regression models

In bootstrapping single-level regression models, we have two basic choices (Mooney & Duvall, 1993; Stine, 1989): bootstrapping cases or bootstrapping residuals. First, we can resample complete cases, as in the previous example. This appears straightforward, but it runs against the assumption that in regression analysis the explanatory variables are viewed as fixed values. This means that in any replication of the study, we expect that the predictor variables have *exactly* the same values and only the residual error and hence the outcome variable will be different. To simulate this situation, we can resample not entire cases, but only the residuals. To bootstrap residuals, we first perform an ordinary multiple regression analysis and estimate the regression coefficients and the residuals. Next, in each bootstrap iteration, the fixed values and regression coefficients are used to produce predicted outcomes, and a bootstrapped set of residuals is randomly added to these predicted outcomes. These bootstrapped responses are then used to estimate the regression coefficients and other parameters of the model.

The choice between bootstrapping cases or residuals depends on the actual design and sampling process. Resampling residuals follows the statistical regression model more accurately. The statistical model behind multiple regression assumes that the predictor variables are fixed by design, and that, if we replicate the study, the explanatory variables have exactly the same values. This can be appropriate if the study is an experiment, with the values of the explanatory variables fixed by the experimental design. However, in much social and behavioral science, the values of the explanatory variables are as much sampled as the responses. In a replication, we do not expect the explanatory variables to have exactly the same values. In this case, resampling cases would be justifiable.

In multilevel regression, bootstrapping cases is more complicated than in ordinary regression models, because it implies bootstrapping units at all available levels. This not only changes the values of the explanatory and outcome variables, but also the multilevel structure: the sample sizes and the way the variance is partitioned over the different levels. For example, imagine sampling cases from the popularity example data, which has 2000 pupils in 100 classes. The class sizes are not all equal, so if we take a bootstrap sample of classes we are likely to have a sample of pupils larger or smaller than 2000. We can adjust by changing the class samples, but then the class sizes are not correct. The redistribution of the variance that also occurs affects all the other estimates. Currently, none of the major software packages supports multilevel casewise bootstrapping. The program MLwiN supports bootstrapping residuals, in two different forms. The first form is the nonparametric bootstrap, in which the multilevel regression estimation is carried out once on the total data set. The regression coefficients from this estimation are used to produce predicted values, and the residuals from this analysis are resampled in the bootstrap iterations.[5] This approach is called the nonparametric bootstrap because it preserves the possibly non-normal distribution of the residuals. An example of using the nonparametric bootstrap to construct confidence intervals for non-normal data is given by Carpenter, Goldstein, and Rasbash (2003). The second approach, the parametric bootstrap, is to simulate the residuals using a multivariate normal distribution. In this approach, the residuals by definition always have a nice normal distribution.

If the observed residuals actually have a normal distribution, the parametric and nonparametric bootstrap are equivalent. In this case, both bootstrap methods can be used to obtain valid estimates of the sampling error in small samples, where the asymptotic standard errors may not be applicable. If the residuals do not have a normal distribution, the nonparametric bootstrap reproduces this irregularity in the bootstrap samples. In theory, this should produce valid standard errors and confidence intervals in these circumstances.

MLwiN contains the bootstrap-based bias correction described earlier, for both the nonparametric and the parametric bootstrap method. The bias correction can be repeated several times, by bootstrapping repeatedly using the corrected parameter estimates. This is called the iterated bootstrap; the complete bootstrap procedure is repeated several times, each time with new model parameters that include the bias corrections suggested by the earlier bootstraps.

Bootstrapping takes the observed data as the sole information about the population, and therefore it is best used with a reasonable level 2 sample size. When we estimate variance components, a minimum of 50 second-level units is recommended for bootstrapping. If the interest is mainly in the fixed effects, which usually have a well-behaving symmetric distribution, we might get away with as few as 10 to 12 units (see Good, 1999, p. 109).

[5] Since the observed residuals are themselves a sample, their variance is not exactly equal to the variance estimated by the maximum likelihood procedure. Therefore, before the residuals are resampled, MLwiN transforms them to make them conform exactly to the estimated variances and covariances at all levels (see Rasbash et al., 2000). Since a model is still needed to estimate the residuals, Carpenter and Bithell (2000) reserve the name 'nonparametric bootstrap' for the cases bootstrap, and call the residuals bootstrap 'semi-parametric'.

### 13.3.3 An example of the multilevel bootstrap

We will use the pupil popularity data introduced in Chapter 2 to illustrate bootstrapping. This is a random component model. By omitting the significant variance for the slope of pupil extraversion, we misspecify the model, and as a result introduce some heteroscedasticity in the second-level residuals. The same model is used in section 13.2 for the robust standard errors, and in section 13.4 on Bayesian estimation.

Using MLwiN, there are several choices in the bootstrapping menu, such as setting the number of iterations for the bootstrap, or the number of iterated bootstraps. One choice is vital: allowing the program to estimate negative variance components. Many programs, including MLwiN, by default set negative variance estimates to zero, because negative variances are impossible. However, setting an offending variance estimate to zero may produce a better estimate, but it will also produce bias. To use bootstrapped estimates to estimate a parameter or establish a confidence interval, we need unbiased estimates in the bootstrap samples.

Table 13.4 presents the results of a parametric bootstrap using three iterated bootstrap runs of $b = 1000$ iterations each. The 95% confidence interval for the bootstrap can be obtained in two different ways: by applying the usual procedure taking the bias-corrected estimate ± 1.96 times the bootstrapped standard error, or by taking the 2.5th and 97.5th percentiles of the bootstrap distribution of the last of the three bootstraps. Especially when bootstrapping parameter estimates that do not have a normal sampling distribution, such as variances, using the percentile method is superior. For the purpose of comparison, Table 13.4 shows both bootstrap intervals: the normal and the percentile method.

*Table 13.4* Comparison of asymptotic and bootstrap results, popularity data

| | ML estimates | | Parametric bootstrap | | |
|---|---|---|---|---|---|
| **Fixed** | Coeff. (s.e.) | 95% CI | Coeff. (s.d.) | 95% CI (normal) | 95% CI (percent) |
| Intercept | 5.07 (.06) | 4.96–5.18 | 5.07 (.06) | 4.96–5.19 | 4.96–5.19 |
| Pup. gndr | 1.25 (.04) | 1.18–1.33 | 1.26 (.04) | 1.18–1.33 | 1.18–1.33 |
| Pup. extrav. | 0.45 (.02) | 0.42–0.49 | 0.46 (.02) | 0.42–0.49 | 0.42–0.49 |
| Tch. exp. | 0.09 (.01) | 0.07–0.11 | 0.09 (.01) | 0.07–0.11 | 0.07–0.11 |
| **Random** | | | | | |
| $\sigma_e^2$ | 0.59 (.02) | 0.55–0.63 | 0.59 (.02) | 0.55–0.63 | 0.56–0.63 |
| $\sigma_{u0}^2$ | 0.29 (.05) | 0.20–0.38 | 0.30 (.05) | 0.20–0.39 | 0.20–0.38 |

The bootstrap results in Table 13.4 are almost identical to the asymptotic estimates. Since we have on purpose omitted a significant variance component for pupil gender, we know that the second-level residuals do not have a normal distribution, and Figure 13.2 confirms this. Therefore, simulating residuals from a normal distribution, as is done in the parametric bootstrap, is not optimal. The nonparametric bootstrap uses the non-normal residuals in the bootstrap samples, and for that reason produces estimates that reflect the underlying distribution better. Table 13.5 shows the results of three iterated nonparametric bootstrap runs of 1000 iterations each.

The bootstrapped results are again very close to the asymptotic estimates, demonstrating that these data are closely following the assumptions for the asymptotic

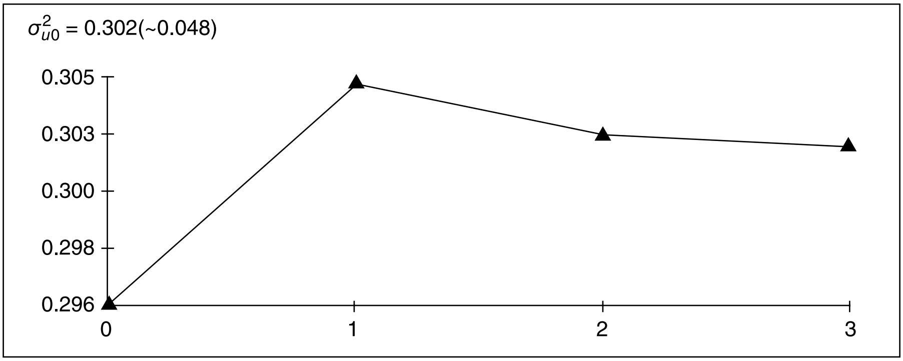


*Figure 13.2* Bias-corrected estimate for $\sigma^2_u$ after three iterated bootstraps.

*Table 13.5* Comparison of asymptotic and iterated bootstrap results, popularity data

| | ML estimates | | Nonparametric bootstrap | | |
|---|---|---|---|---|---|
| **Fixed** | Coeff. (s.e.) | 95% CI | Coeff. (s.d.) | 95% CI (normal) | 95% CI (percent) |
| Intercept | 5.07 (.06) | 4.96–5.18 | 5.08 (.06) | 4.97–5.19 | 4.96–5.19 |
| Pup. gndr | 1.25 (.04) | 1.18–1.33 | 1.26 (.04) | 1.18–1.33 | 1.18–1.33 |
| Pup. extrav. | 0.45 (.02) | 0.42–0.49 | 0.46 (.02) | 0.42–0.49 | 0.42–0.49 |
| Tch. exp. | 0.09 (.01) | 0.07–0.11 | 0.09 (.01) | 0.07–0.11 | 0.07–0.11 |
| **Random** | | | | | |
| $\sigma^2_e$ | 0.59 (.02) | 0.55–0.63 | 0.59 (.02) | 0.55–0.63 | 0.56–0.63 |
| $\sigma^2_{u0}$ | 0.29 (.05) | 0.20–0.38 | 0.30 (.05) | 0.21–0.40 | 0.21–0.40 |

estimates. The bias-corrected estimates are close to the asymptotic estimates, indicating that there is no important bias in the asymptotic estimates. If there is a distinct difference between the asymptotic and the bias-corrected parameter estimates, the estimates of the successive iterated bootstraps should be monitored, to check that the series of iterated bootstraps has converged with sufficient accuracy. By way of example, Figure 13.2 shows the trend for the asymptotic and the bias-corrected estimate for the class-level variance component $\sigma^2_{u0}$ in a series of three iterated bootstraps.

There is a very small bias correction visible in the first bootstrap iteration, and the second and third bootstraps do not change the estimate much. Therefore, we conclude that the iterated bootstrap has converged. The difference between the asymptotic estimate of 0.296 and the final bias-corrected estimate of 0.304 is of course trivial. It is as an indication of a real, but very small and therefore in practice negligible, negative bias in the second-level variance estimate.

## 13.4 BAYESIAN ESTIMATION METHODS

Statistics is about uncertainty. We estimate unknown population parameters by statistics, calculated in a sample. In classical statistical inference, we express our uncertainty about how well an observed statistic estimates the unknown population parameter by examining its sampling distribution over an infinite number of possible samples. Since we generally have only one sample, the sampling distribution is based on a mathematical sampling model. An alternative basis is bootstrapping, discussed in the previous section, which simulates the sampling process. The sampling variance, or rather its square root, the standard error, is used for significance testing and establishing confidence intervals.

In Bayesian statistics, we express the uncertainty about the population value of a model parameter by assigning to it a probability distribution of possible values. This probability distribution is called the *prior* distribution, because it is specified independently from the data. The Bayesian approach is fundamentally different from classical statistics (Barnett, 1999). In classical statistics, the population parameter has one specific value, only we happen not to know it. In Bayesian statistics, we consider a probability distribution of possible values for the unknown population parameter. After we have collected our data, this *prior distribution* is combined with the likelihood of the data to produce a *posterior* distribution, which describes our uncertainty about the population values after observing our data. Typically, the variance of the posterior distribution is smaller than the variance of the prior distribution, which means that observing the data has reduced our uncertainty about the possible population values.

In Bayesian statistics, each unknown parameter in the model must have an associated probability distribution. For the prior distribution, we have a fundamental choice between using an informative prior or an uninformative prior. An informative prior is a peaked distribution with a small variance, which expresses a strong belief about the unknown population parameter. An informative prior will, of course, strongly influence the posterior distribution, and hence our conclusions. For this reason, many statisticians prefer an uninformative or diffuse prior, which has very little influence on the conclusions, and only serves to produce the posterior. An example of an uninformative prior is the uniform distribution, which simply states that the unknown parameter value is between minus and plus infinity, with all values equally likely. Another example of an uninformative prior is a very flat normal distribution. Sometimes such a prior is called an ignorance prior, to indicate that we know nothing about the unknown parameter. However, this is not accurate, since total ignorance does not exist, at least not in Bayesian statistics. All priors add some information to the data, but diffuse priors add very little information, and therefore do not have much influence on the posterior. One way to express the information added to the data is to view the prior as a certain number of hypothetical cases, which are added to the data set. Typically, an ignorance prior corresponds to fewer than 10 of such hypothetical cases, which will have little influence unless the observed sample is very small.

If the posterior distribution has a mathematically simple form, such as a normal distribution, we can use the known characteristics of this distribution to calculate a point estimate and a confidence interval for the population parameter. In the case of a normal distribution, we could choose the mean as the point estimate, and base a confidence interval on the standard deviation of the posterior distribution. However, when Bayesian methods are applied to complex multivariate models, the posterior is generally a multivariate distribution with a complicated shape, which makes it difficult to use mathematical means to establish confidence intervals. When the posterior distribution is difficult to describe mathematically, it is approximated using Markov chain Monte Carlo simulation procedures. Markov chain Monte Carlo (MCMC) procedures are simulation techniques that generate random samples from a complex posterior distribution. By producing a large number of random draws from the posterior distribution, we can closely approximate its true shape. The simulated posterior distribution is then used to compute a point estimate and a confidence interval (for an accessible introduction to Bayesian MCMC methods see Casella & George, 1992; Gelman & Hill, 2007; Smith & Gelfland, 1992). Typically, the marginal (univariate) distribution of each parameter is used. The mode of the marginal posterior distribution is an attractive point estimate of the unknown parameter, because it is the most likely value, and therefore the Bayesian equivalent of the maximum likelihood estimator. Since the mode is more difficult to determine than the mean, the mean of the posterior distribution is also often used. In skewed posterior distributions, the median is an attractive choice. The confidence interval generally uses the $100 - \frac{1}{2}\alpha$ limits around the point estimate. In the Bayesian terminology, this is referred to as the 100 – α *central credibility interval.*

Bayesian methods have some advantages over classical methods. To begin, in contrast to the asymptotic maximum likelihood method, they are valid in small samples. Given the correct probability distribution, the estimates are always proper, which solves the problem of negative variance estimates. Finally, since the random draws are taken from the correct distribution, there is no assumption of normality when variances are estimated.

### 13.4.1 Simulating the posterior distribution

Different simulation methods are used to generate draws from the posterior distribution. Most methods use Markov chain Monte Carlo sampling. Given a set of initial values from a specific multivariate distribution, MCMC procedures generate a new random draw from the same distribution. Suppose that $Z^{(1)}$ is a draw from a target distribution $f(Z)$. Using MCMC methods, we generate a series of new draws: $Z^{(1)}\rightarrow Z^{(2)}\rightarrow \ldots \rightarrow Z^{(t)}$. MCMC methods are attractive because, even if $Z^{(1)}$ is not from the target distribution $f(Z)$, if $t$ is sufficiently large, in the end $Z^{(t)}$ is a draw from the target distribution $f(Z)$. Having good initial values for $Z^{(1)}$ helps, because it speeds up the convergence on the target distribution, so the classical maximum likelihood estimates are often used as initial values for $Z^{(1)}$.

The number of iterations $t$ needed before the target distribution is reached is referred to as the ‘burn-in’ period of the MCMC algorithm. It is important that the burn-in is complete. To check if enough iterations of the algorithm have passed to converge on the target distribution, several diagnostics are used. A useful diagnostic is a graph of the successive values produced by the algorithm. A different procedure is to start the MCMC procedure several times with widely different initial values. If essentially identical distributions are obtained after $t$ iterations, we decide that $t$ has been large enough to converge on the target distribution (Gelman & Rubin, 1992).

An additional issue in MCMC methods is that successive draws are dependent. Depending on the distribution and the amount of information in the data, they can be strongly correlated. Logically, we would prefer independent draws to use as simulated draws from the posterior distribution. One way to reach independence is to throw away a number of successive estimates before a new draw is used for estimation. This process is called *thinning*. To decide how many iterations must be thrown away between two successive draws, it is useful to inspect the autocorrelations between successive draws. If the autocorrelations are high, we must throw away many estimates. Alternatively, since each draw still gives some information, we may keep all of them, but use an extremely large number of draws. Typically, the number of MCMC iterations is much higher than the number of bootstrap samples. Using 10,000 or more MCMC iterations is common.

### 13.4.2 Bayesian estimation using MLwiN: The estrone data

User-friendly software for MCMC estimation is still rare; most Bayesian modeling is executed using special software (see Gelman & Hill, 2007, for multilevel applications of MCMC estimation). A very general but difficult program is *BUGS*, an acronym for Bayesian inference using Gibbs sampling (Spiegelhalter, 1994). At the time of writing, the only multilevel software products that include Bayesian methods are MLwiN and R (see Gelman and Hill, 2007, for a thorough description of Bayesian regression modeling in R). It is instructive to use the MCMC method in MLwiN to analyze the estrone data. The issues of determining the number of MCMC iterations, monitoring the convergence of the MCMC algorithm using plots and special statistics, and settling on estimates and confidence intervals are not dependent on the specific software used. Since the data set is very small (16 multiple measures on five women), asymptotic maximum likelihood does not work well for these data, and Bayesian methods may do better.

By default, MLwiN uses non-informative priors. To start, we use the default burn-in period of 500 iterations, and 5000 subsequent iterations for the MCMC chain. MLwiN produces a variety of plots and statistics to judge whether these quantities are sufficient. Figure 13.3 shows a plot of the first 500 burn-in estimates for the intercept $\beta_0$. It is clear from this figure that the series of estimates shows a slight trend, and that this trend has not yet flattened out after 500 initial iterations. This suggests that we need more than 500 initial iterations for the burn-in.

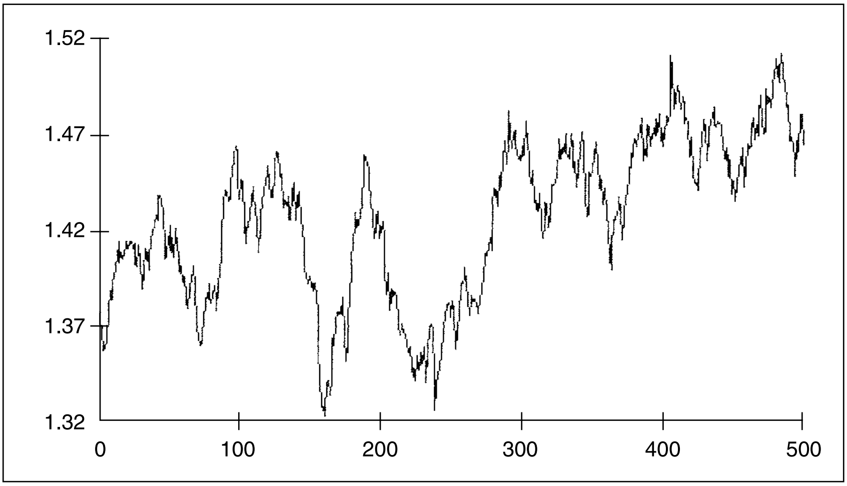


*Figure 13.3* Plot of 500 burn-in estimates of the intercept $\beta_0$, estrone data.

Figure 13.3 shows that successive parameter estimates tend to be similar. This indicates that the successive draws are not independent, which results in autocorrelation of the series of estimates. MLwiN can produce an *autocorrelation plot*, called the autocorrelation function or ACF in the output. This plot shows the correlation between two MCMC draws that are separated by 1, 2, 3, and so on, MCMC iterations. Figure 13.4 presents the autocorrelation plot for $\beta_0$.

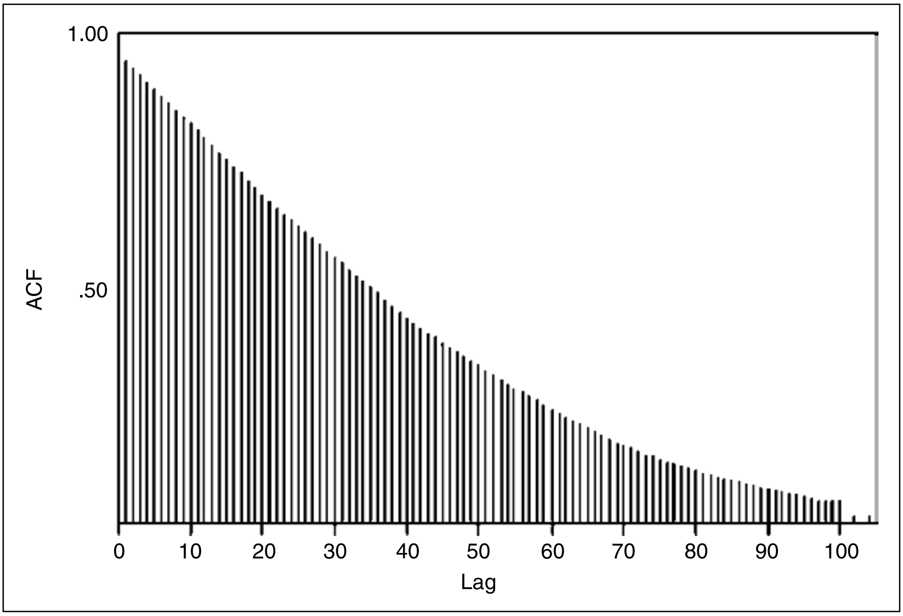


*Figure 13.4* Autocorrelations for the intercept estimates $\beta_0$, estrone data.

Figure 13.4 shows strong autocorrelations between successive draws. When draws are more than 40 iterations apart, the autocorrelation is below .50, but still considerable. Even after 100 iterations, the autocorrelation is still not close to zero. When we inspect the autocorrelation plots for the variance estimates $\sigma_e^2$ and $\sigma_{u0}^2$, we observe similar problems, but not as severe as the problems in generating estimates for the intercept, since the autocorrelation is lower.

Based on the plots in Figure 13.3 and 13.4, we decide to use a burn-in period of 5000. Figure 13.5 shows the first 5000 estimates for the intercept. The plot of these 5000 estimates suggests that the series of estimates converges on the correct distribution after about 2000 iterations. Of course, the autocorrelations are still high. To

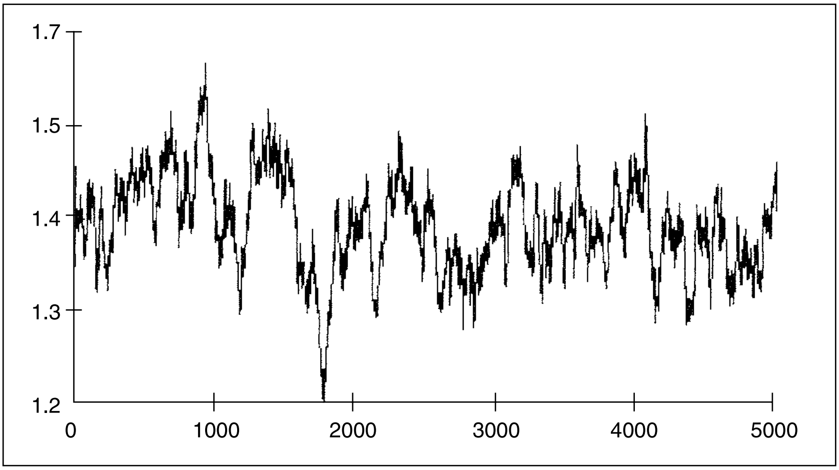


*Figure 13.5* Plot of first 5000 estimates of the intercept $\beta_0$, estrone data.

obtain usable estimates, a burn-in of 5000 seems sufficient. To deal with the high autocorrelations, we call for 500,000 MCMC iterations after the burn-in, and use a *thinning factor* of 100. That is, to reduce memory demands the program stores each 100th set of MCMC estimates, and discards the other estimates. This will give us 5000 MCMC estimates, each 100 MCMC iterations apart, which reduces the autocorrelations considerably.

Figure 13.6 shows the last 500 estimates for the intercept. Note that these are the thinned estimates taken at each 100th iteration. The plot looks horizontal and stable, but it still shows evidence of considerable autocorrelation.

Figure 13.7 shows some other diagnostic plots offered by MLwiN. The plot of the total thinned chain of 5000 estimates looks quite stable, and the distribution of the generated values is nearly normal. There is still some autocorrelation. MLwiN produces several diagnostics to evaluate the accuracy of the MCMC estimates. The Raftery-Lewis (Raftery & Lewis, 1992) diagnostic is an estimate of the number of MCMC iterations needed to be 95% confident that the 2.5th and 97.5th percentiles are estimated with an error smaller than 0.005. Typically, the Raftery-Lewis diagnostics (one for each boundary) suggest a very large number of iterations are required to achieve this level of accuracy. For our example data, they indicate that we need to carry out about 1,000,000 MCMC iterations. This is an enormous number, especially since we are already using a thinning factor of 100. The Brooks-Draper diagnostic indicates the length of chain required to produce an estimate of the mean accurate to *n*

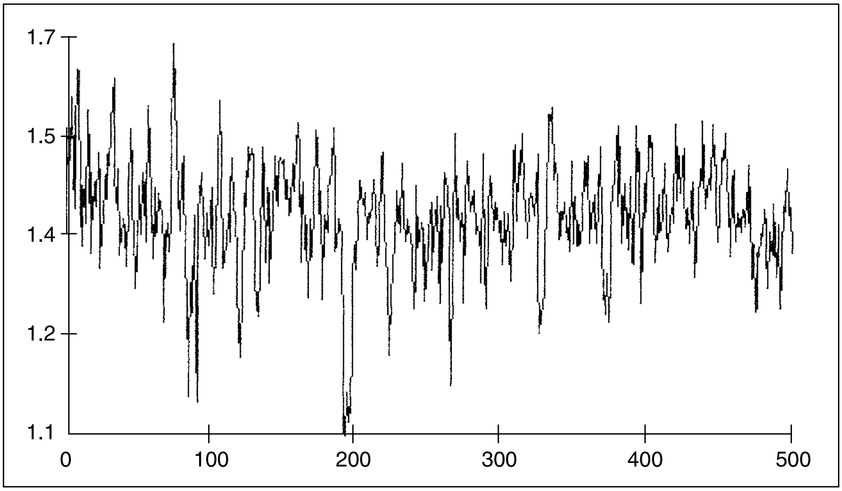


*Figure 13.6* Plot of last 500 estimates of the intercept $\beta_0$, estrone data.

significant figures. In the estrone example, the Brooks-Draper indicates that we need 39 iterations to estimate the mean within two significant decimals (this is of course after the thinning factor). MLwiN also reports an effective sample size (ESS) of 1325 for our estimates, which means that the autocorrelation between the draws (which are thinned by a factor of 100) reduces our chain of 5000 iterations to an equivalent of 1325 independent draws.

Given the normal distribution of the intercept estimates, we can use the mode of 1.42 as a point estimate. The standard deviation of the MCMC estimates is 0.083, which we can use as a standard error in the usual way. This produces a Bayesian 95% credibility interval of 1.26–1.58. The Bayesian central 95% credibility interval determined from the 2.5th and 97.5th percentiles of the 5000 observed estimates is 1.24–1.58, which is very close. The maximum likelihood point estimate is 1.42, with a standard error of .06, and a 95% confidence interval of 1.30–1.54. Since maximum likelihood is applied here in a very small sample, the MCMC confidence intervals are likely to be more realistic.

In this example, we are mostly interested in the between-subject variance estimate $\sigma^2_{u0}$. Figure 13.8 presents the plot of the last 500 estimates for $\sigma^2_{u0}$. The plot of the variance appears stable; most estimates are close to zero, with a few spikes that indicate an occasional large estimate. Given the small sample of subjects, this is normal.

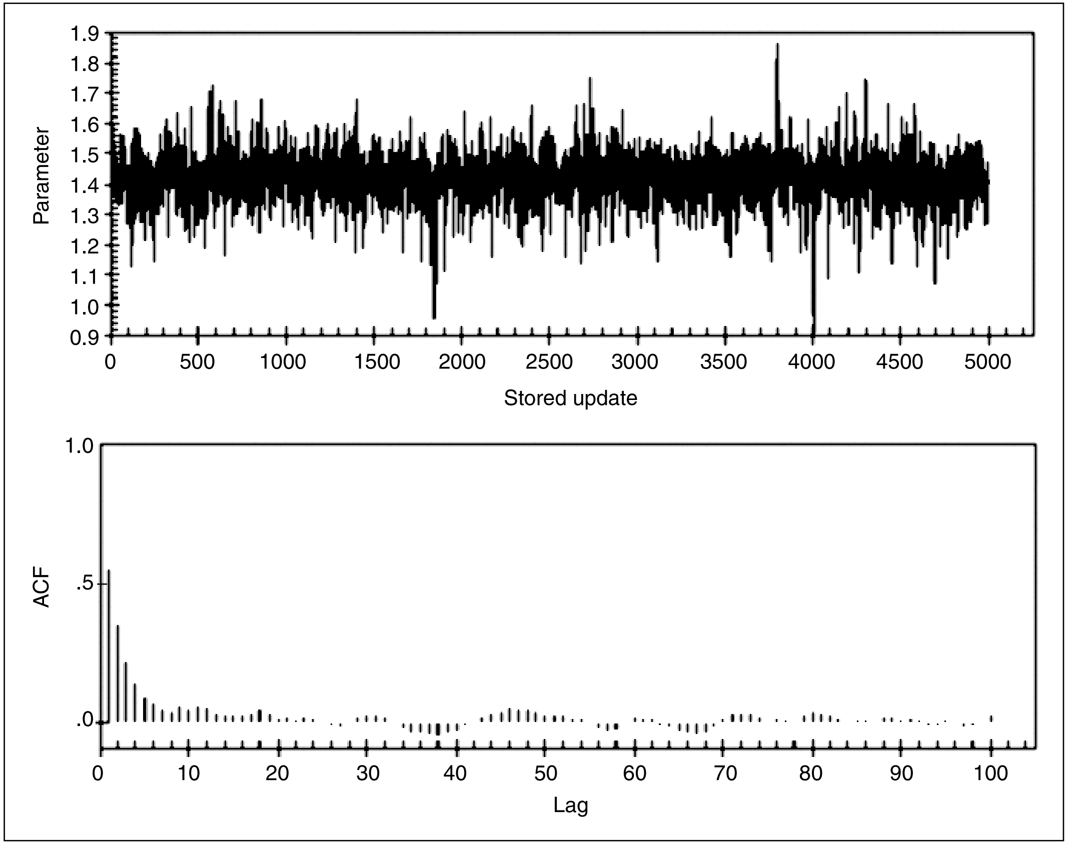


*Figure 13.7* Diagnostic plots for 5000 estimates of the intercept $\beta_0$, estrone data.

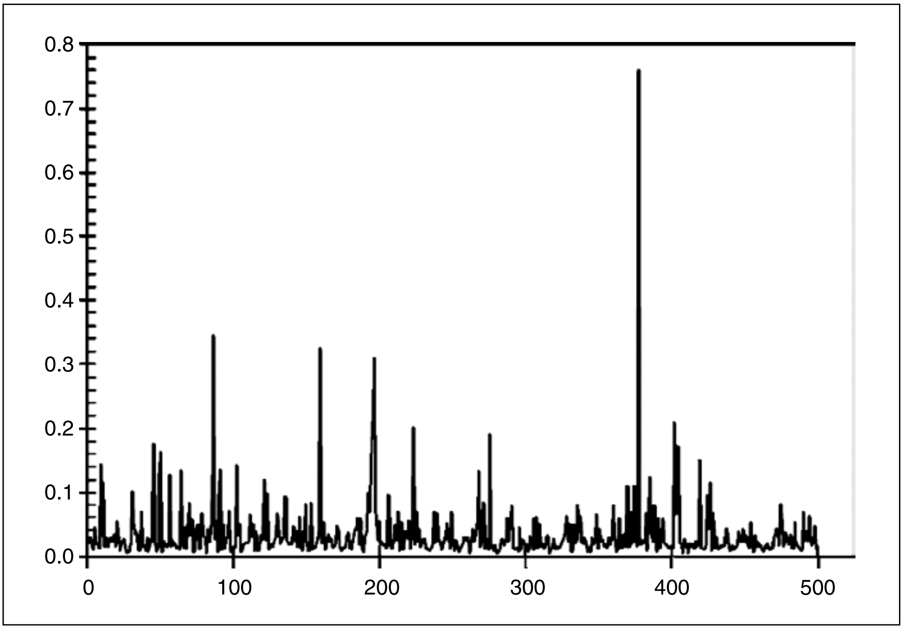


*Figure 13.8* Plot of last 500 estimates of the variance $\sigma_{u0}^2$, estrone data.

Figure 13.9 presents the same diagnostic plots for all 5000 stored MCMC estimates for the estrone data. The diagnostic plots for the variance in Figure 13.9 highlight an important feature of the Bayesian estimation method used in MCMC: it always produces proper estimates. Thus, it will never produce a negative variance estimate. When we generate values for a variance that is actually zero, unbiased estimates should either be all equal to zero, or vary around this zero value, with approximately half of the estimates being positive and the other half negative.

Since MCMC methods will never produce negative variance estimates, they have a positive bias. As a result, *no* central credibility interval for a variance component will *ever* include the value zero. For instance, in the estrone example, the central 95% interval for the variance estimates is 0.01–0.14. If the variance term $\sigma_{u0}^2$ indeed belongs in the model, the 0.01–0.14 interval is a reasonable 95% confidence interval, although the Raftery-Lewis statistic again indicates that we should use many more MCMC iterations for accuracy. However, since the value of the variance term is very small, we may well conclude that the between-subject variance term is not significant, and can be omitted from the model. The fact that the value zero is outside the central 95% interval is no evidence that the variance term is significant, because when variances are estimated, the value zero will *always* lie outside the central 95% interval. To carry out a

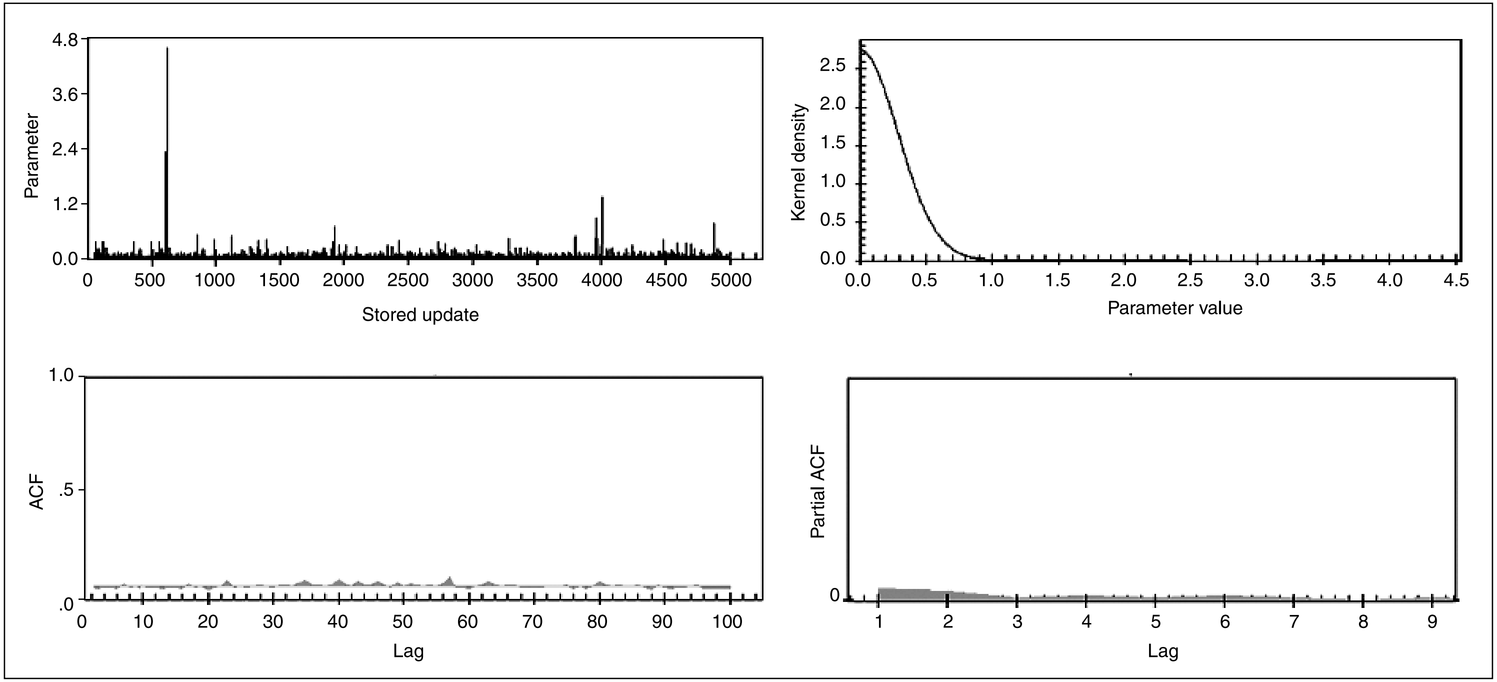


*Figure 13.9* Diagnostic plots for 5000 estimates of the variance $\sigma_0^2$, estrone data.

Bayesian significance test to determine if $\sigma_{u0}^2$ belongs in the model at all, we need different methods, which are not implemented in MLwiN. For an example of Bayesian *p*-values see Hoijtink (2000). Bayesian statisticians prefer to use model selection methods, comparing models with and without the subject-level variance term, to decide which model is best. The BIC (Bayesian information criterion) fit index discussed in Chapter 3 is related to these Bayesian model selection procedures. For an introduction including multilevel methods see Gelman and Hill (2007); for a discussion in the context of MLwiN see Browne (2005).

### 13.4.3 An example of Bayesian estimation using MLwiN: The popularity data

The estrone example illustrates that Bayesian methods do not solve all problems. The MCMC method does not deal very well with the estrone data set. The number of MCMC iterations that are required is very large, and the autocorrelations are high, especially in estimating the mean. This indicates that the data contain very little information about the parameter estimates, which is not surprising given the small sample size.

To illustrate Bayesian MCMC methods on a less extreme data set, they are also applied to the popularity data set, which consists of 2000 pupils in 100 classes. As in the section on bootstrapping, the model is a variance component model, on purpose omitting the significant variance term for pupil extraversion and the cross-level interaction of pupil extraversion and teacher experience. To facilitate estimation, all predictor variables are grand mean-centered. Using 500 iterations for the burn-in and a chain of 5000 for the estimates, MLwiN produces the plots shown in Figure 13.10.

Figure 13.10 shows the plots of the last 500 estimates. All plots look well-behaved, meaning that no marked overall trend is visible, and the generating process appears stable. Both the intercept ($\beta_0$) and the regression coefficient of teacher experience ($\beta_3$) show rather strong autocorrelations (estimates close to each other are similar). All other plots look fine: chaotic, without obvious patterns or trends.

To examine the autocorrelation, we inspect the diagnostics for the slope of teacher experience. Figure 13.11 looks acceptable, but there is indeed a fair amount of autocorrelation, which disappears after about 40 MCMC iterations.

In addition to the plots, the convergence of the MCMC chain can also be studied by starting it with different initial values, and inspecting if, and after how many MCMC iterations, the different chains converge on the same distribution. For instance, we may replace the customary maximum likelihood starting values by different values such as: intercept is 0 (FIML: 5.07); slope for pupil gender is 0.5 (FIML: 1.25), slope for pupil extraversion is 2 (FIML: 0.45), slope for teacher experience is 0.1 (FIML:

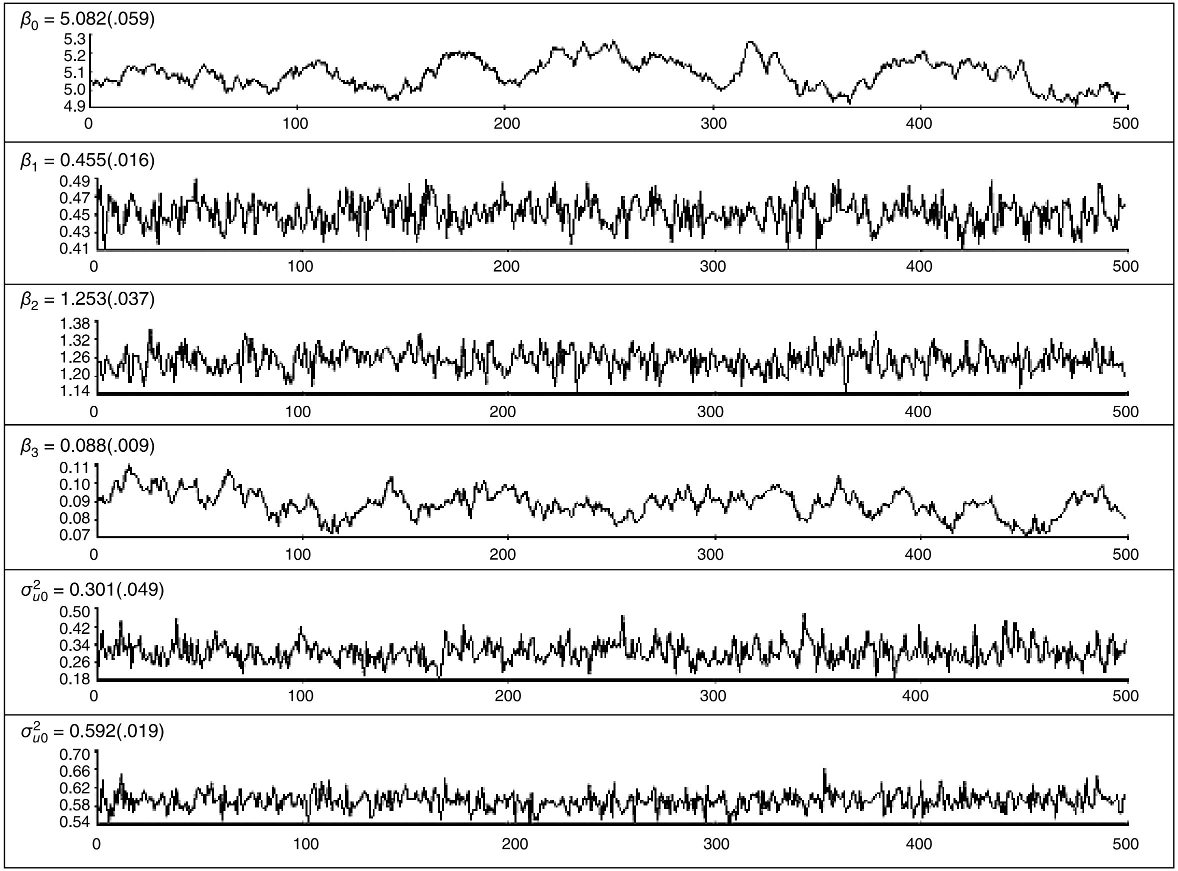


*Figure 13.10* Plot of last 500 estimates of all parameters, popularity data.

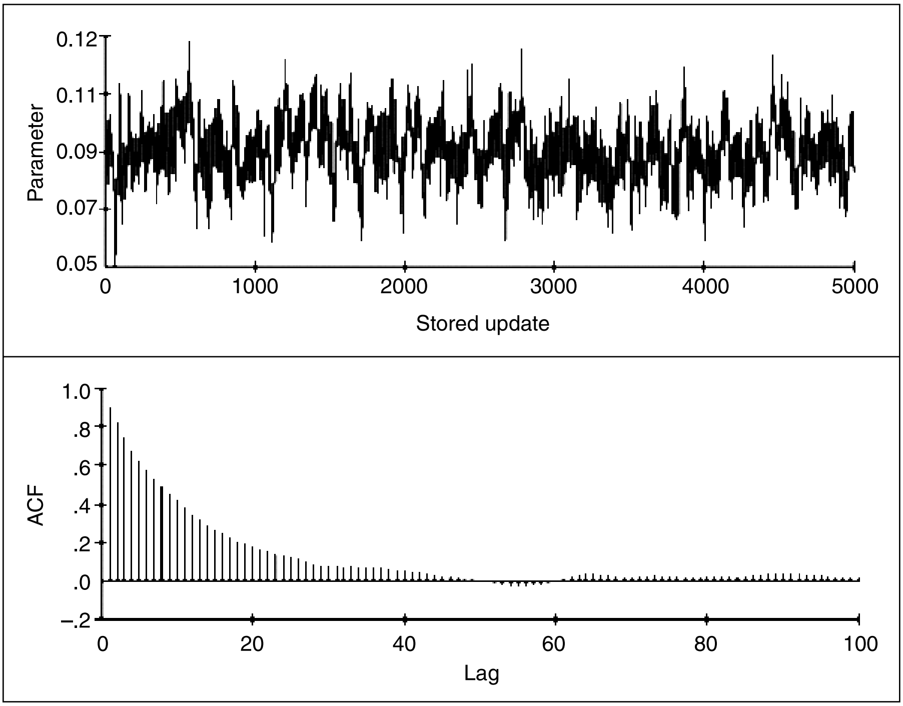


*Figure 13.11* Diagnostic plots for 5000 estimates of the teacher experience slope.

0.09), variance class level is 0.2 (FIML: 0.29), and pupil level is 0.8 (FIML: 0.59). These initial values are reasonable, but deviate noticeably from the maximum likelihood estimates. If we monitor the burn-in process that starts with these initial estimates, we find the plots given in Figure 13.12.

Figure 13.12 nicely illustrates what happens if we start the MCMC algorithm with poor starting values in the presence of an informative data set. It prolongs the burn-in period. In this case, it is clear that after about 200–400 iterations, the successive MCMC estimates have stabilized. Only the intercept term $\beta_0$ might need a longer burn-in. This result reassures us that the MCMC chain is converging on the correct distribution. If the data contain little information on the model parameters (which is another way of stating that the model is too complex for the available data), the MCMC chains with different starting values would not converge, or would converge extremely slowly. The appropriate action would be to simplify the model.

For the final MCMC analysis, the analysis is repeated, with the burn-in set at 1000, and a MCMC chain of 50,000 iterations, with a thinning factor of 10. In this analysis, all diagnostics look fine. Table 13.6 presents the results, together with the

*Table 13.6* Comparison of asymptotic and Bayesian results, popularity data

| | ML estimates | | MCMC estimates | |
|---|---|---|---|---|
| **Fixed** | Coeff. (s.e.) | 95% CI | Coeff. (s.d.) | 95% CI |
| Intercept | 5.07 (.06) | 4.96–5.18 | 5.07 (.06) | 4.96–5.19 |
| Pup. gender | 1.25 (.04) | 1.18–1.33 | 1.25 (.04) | 1.18–1.33 |
| Pup. extrav. | 0.45 (.02) | 0.42–0.49 | 0.45 (.02) | 0.42–0.49 |
| Teacher exp. | 0.09 (.01) | 0.07–0.11 | 0.09 (.01) | 0.07–0.11 |
| **Random** | | | | |
| $\sigma_e^2$ | 0.59 (.02) | 0.55–0.63 | 0.59 (.02) | 0.56–0.63 |
| $\sigma_{u0}^2$ | 0.29 (.05) | 0.20–0.38 | 0.30 (.05) | 0.22–0.41 |

asymptotic estimates. The maximum likelihood estimates are based on FML, and the 95% confidence interval is based on the standard normal approximation. For the Bayesian results, the posterior mode is used as the point estimate, because with normal data this is equivalent to the maximum likelihood estimate. The Bayesian central 95% credibility interval is based on the 2.5th and 97.5th percentile points of the posterior distribution.

The Bayesian estimates in Table 13.6 are very close to the maximum likelihood estimates. Only the 95% central confidence interval for the class-level variance is somewhat different. Just as in the bootstrap example, the observed interval is not symmetric around the modal estimate, which reflects the non-normal distribution of the variance parameter.

### 13.4.4 Some remarks on Bayesian estimation methods

An advantage of the Bayesian approach is that when the posterior distribution is simulated, the uncertainty of the parameter estimates is taken into account. So, the uncertainty in the parameter estimates for the fixed part is taken into account in the estimates for the random part. Moreover, simulating a large sample from the posterior distribution is useful because it provides not only point estimates (i.e., posterior mode or mean) of the unknown parameters, but also confidence intervals that do not rely on the assumption of normality for the posterior distribution. As a result, confidence intervals are also accurate for small samples (Tanner & Wong, 1987). The estrone example illustrates that Bayesian methods do not solve all problems. Clearly, the MCMC method does not deal very well with this extremely small data set. The number of MCMC iterations that are required are very large, and the autocorrelation is

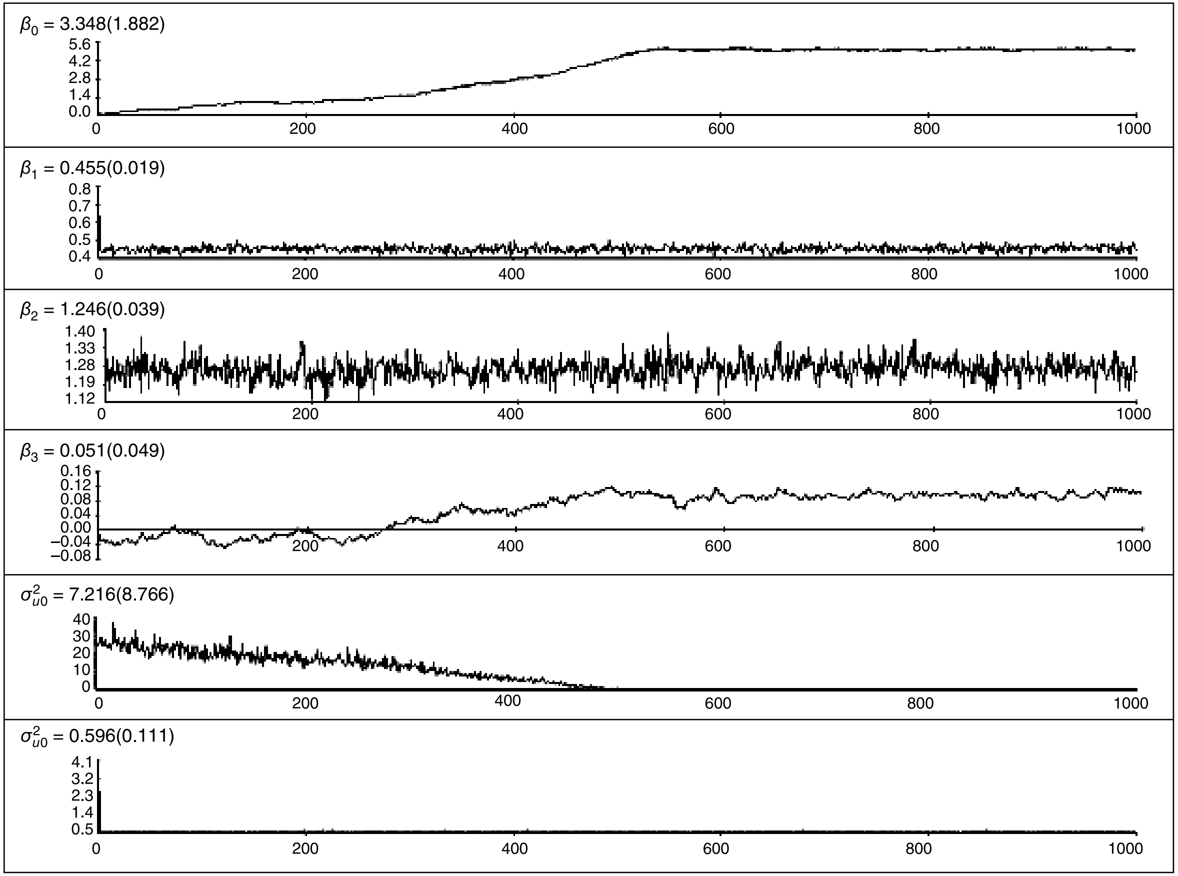


*Figure 13.12* Plot of first 1000 estimates, deviant initial values.

relatively high, especially in estimating the mean. This indicates that the data contain very little information about the parameter estimates, which is not surprising given the small sample size.

In the popularity example, the Bayesian method performs much better. But even in this well-behaved data set, the analyst must inspect the output carefully for indications of non-convergence or other problems. In the final analysis, with the burn-in set at 1000, and a MCMC chain of 50,000 iterations, with a thinning factor of 10, the Raftery-Lewis statistic still indicates that more iterations are needed. However, this statistic is rather conservative. The Raftery-Lewis statistic is a lower-bound estimate (Raftery & Lewis, 1992) that often indicates huge numbers of MCMC iterations. It estimates the number of MCMC iterations needed to be 95% confident that the 2.5th and 97.5th percentiles are estimated with an error smaller than 0.005. In essence, it requires that we are 95% confident that the end points of the 95% confidence interval are correct in two decimal places. If several chains of far smaller numbers of iterations converge on the same confidence interval estimates, we may conclude that the smaller number is sufficient. In fact, if we analyze the pupil popularity data using the standard 500 iterations for the burn-in with 5000 consecutive iterations (no thinning) for monitoring, we find almost the same results.

Bayesian estimation methods involve all the usual assumptions. They are *not* nonparametric, like the nonparametric bootstrap. They use different methods to find point estimates and assess sampling variance, and they can be used in situations where asymptotic maximum likelihood methods are problematical.

As indicated earlier, all priors add some information to the data. As a result, Bayesian estimates are biased. When the sample size is reasonable, the amount of bias is small. This bias is acceptable, because Bayesian methods promise more precision, and better estimates of the sampling variance. Simulation research (Browne, 1998; Browne & Draper, 2000) confirms this, especially when we are dealing with nonlinear models.

In this section, all priors used are non-informative. This is useful, because with normal data and large datasets, the resulting estimates are equivalent to maximum likelihood estimates. In addition, most analysts would be cautious about adding prior information to the data, because this could be interpreted as manipulating the outcome. However, sometimes we do have valid prior information. For instance, when we use a normed intelligence test as an outcome variable, we know that its mean is in the neighborhood of 100, with a standard deviation of about 15. If we use a 10-point scale, we know that the mean must be somewhere between zero and 10, and the variance cannot be larger than 25. So, using a prior for the variance that implies that the variance can be any number between zero and infinity appears to be wasting real information. Novick and Jackson (1974) suggest in their excellent introduction to Bayesian methods that in a scientific debate it can be constructive to define two different priors that correspond to two different hypotheses about a phenomenon. After data are collected and analyzed, the two posterior distributions should be more similar than the two prior distributions, indicating that observing the data allows the conclusions to converge. An interesting example is given by Meyer (1964), and discussed in detail by Novick and Jackson (1974, pp. 177–182).

It is clear that Bayesian MCMC methods, like bootstrapping, are computationally intensive. However, given modern computer equipment and reasonably sized models, they are well within present computer capabilities. The techniques presented in this section all use MLwiN (version 2.10). However, the issues that are addressed, such as deciding on the length of the burn-in and monitoring the MCMC chains, are general. These decisions must be made by the analyst, and should be based on careful inspection of relevant plots and diagnostics. The number of iterations is usually much larger than in bootstrapping. However, since MCMC methods are based on generating estimates, and bootstrap methods on generating or shuffling around raw data, MCMC methods are often faster. As in bootstrapping, the rule is that, statistically speaking, there is no such thing as too many iterations. It may be worthwhile to repeat the analysis of the final model(s) using the very long MCMC chains typically indicated by the Raftery-Lewis statistic.

# 14

# Multilevel Factor Models

The models described in the previous chapters are all multilevel variants of the conventional multiple regression model. This is not as restrictive as it may seem, since the multiple regression model is very flexible and can be used in many different applications (for detailed examples see Cohen & Cohen, 1983). Still, there are models that cannot be analyzed with multiple regression, notably factor analysis and path analysis models.

A general approach that includes both factor and path analysis is *structural equation modeling*, or *SEM*. The interest in structural equation modeling is generally on theoretical constructs, which are represented by latent factors. A factor model, which is often called the measurement model, specifies how the latent factors are measured by the observed variables. The relationships between the theoretical constructs are represented by regression or path coefficients between the factors. The structural equation model implies a structure for the covariances between the observed variables, which explains the alternative name of *covariance structure analysis*. However, the model can be extended to include means of observed variables or factors in the model, which makes 'covariance structure analysis' a less accurate name. Many researchers will simply think of these models as 'LISREL models', which is also less accurate. LISREL is an abbreviation of LInear Structural RELations, and the name used by Jöreskog for one of the first and most popular SEM programs. Nowadays structural equation models need not be linear, and the possibilities of SEM extend well beyond the original LISREL program. Marcoulides and Schumacker (2001), for instance, discuss the possibility of fitting nonlinear curves and interaction terms in SEM.

Structural equation modeling is a general and convenient framework for statistical analysis that includes as special cases several traditional multivariate procedures, such as factor analysis, multiple regression analysis, discriminant analysis, and canonical correlation. Structural equation models are often visualized by a graphical *path diagram*. The statistical model is usually represented in a set of matrix equations. In the early 1970s, when this technique was first introduced in social and behavioral research, the software usually required setups that specified the model in terms of these matrices. Thus, researchers had to distill the matrix representation from the path diagram, and provide the software with a series of matrices for the different sets of parameters, such as factor loadings and regression coefficients. A recent development is software that allows the researchers to specify the model directly as a path diagram. Since the focus in this chapter is on structural equation models for multilevel data, and not on structural equation modeling itself, the models will generally be introduced using path diagrams.

Structural equation modeling has its roots in path analysis, which was invented by the geneticist Sewall Wright (Wright, 1921). It is customary to start a SEM analysis by drawing a path diagram. A path diagram consists of boxes and circles that are connected by arrows. In Wright's notation, observed (or measured) variables are represented by a rectangle or a square box, and latent (or unmeasured) factors by a circle or an ellipse. Single-headed arrows or 'paths' are used to define hypothesized causal relationships in the model, with the variable at the tail of the arrow being the cause of the variable at the point. Double-headed arrows indicate covariances or correlations, without a causal interpretation. Statistically, the single-headed arrows or paths represent regression coefficients, and double-headed arrows represent covariances.

Often a distinction is made between the measurement model and the structural model. The measurement model, which is a confirmatory factor model, specifies how the latent factors are related to the observed variables. The structural model specifies the relationships between the latent factors. In this chapter, I discuss multilevel factor analysis, and introduce the techniques currently available for estimating multilevel factor models. Multilevel path models, which are structural models that may or may not include latent factors, are discussed in Chapter 15. For a general introduction to SEM, I refer to the introductory article by Hox and Bechger (1998) or introductory books such as Loehlin (2004) and Schumacker and Lomax (2004). A statistical treatment is presented by Bollen (1989). An interesting collection of introductory articles focusing on SEM models for multigroup and longitudinal data is found in Little, Schnabel, and Baumert (2000).

Structural equation models are often specified as models for the means and the covariance matrix of multivariate normal data. The model is then:

$$\mathbf{y}_i = \boldsymbol{\mu} + \Lambda\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i, \tag{14.1}$$

which states that the observed variables $\mathbf{y}$ are predicted by a regression equation involving an intercept $\boldsymbol{\mu}$ and the regression coefficients or loadings in factor matrix $\boldsymbol{\Lambda}$ multiplied by the unobserved factor scores $\boldsymbol{\eta}$ plus a residual measurement error $\boldsymbol{\varepsilon}$. This can then be expressed as a model for the covariance matrix $\boldsymbol{\Sigma}$ by:

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda\Phi\Lambda}' + \boldsymbol{\Theta}, \tag{14.2}$$

where the covariance matrix $\boldsymbol{\Sigma}$ is expressed as a function of the factor matrix $\boldsymbol{\Lambda}$, the matrix of covariances between factors $\Phi$, and residual measurement errors in $\Theta$.

This chapter discusses two different appoaches to multilevel SEM. The first approach, described by Rabe-Hesketh, Skrondal, and Zheng (2007) as the 'within and between formulation', focuses on determining separate estimates for the within (subject-level) covariance matrix and the between (group-level) covariance matrix. These are then modeled separately or simultaneously by a subject-level (within) factor model and a group-level (between) factor model, analogous to the single-level equation given in 14.2. This works well, but has limitations, which will be discussed in the next section (14.1) that describes this approach in more detail. The second approach models the observed multilevel data directly with a model that includes variables at each available level and accommodates group-level variation of intercepts and slopes. It is the most accurate and versatile approach, but in some circumstances computationally demanding. It is also at the moment not widely available in standard multilevel or SEM software. This approach is described in section 14.2.

## 14.1 THE WITHIN AND BETWEEN APPROACH

The within and between approach is based on an analysis of the subject-level and the group-level covariance matrix. This is in turn based on a decomposition of the variables to the available levels, which is discussed in the next section.

### 14.1.1 Decomposing multilevel variables

Multilevel structural models assume that we have a population of individuals that are divided into groups. The individual data are collected in a $p$-variate vector $\mathbf{Y}_{ij}$ (subscript $i$ for individuals, $j$ for groups). Cronbach and Webb (1979) have proposed decomposing the individual data $\mathbf{Y}_{ij}$ into a between-groups component $\mathbf{Y}_B = \bar{\mathbf{Y}}_j$, and a within-groups component $\mathbf{Y}_w = \mathbf{Y}_{ij} - \bar{\mathbf{Y}}_j$. In other words, for each individual we replace the observed total score $\mathbf{Y}_T = \mathbf{Y}_{ij}$ by its components: the group component $\mathbf{Y}_B$ (the disaggregated group mean) and the individual component $\mathbf{Y}_W$ (the individual deviation from the group mean.) These two components have the attractive property that they are orthogonal and additive (see Searle, Casella, & McCulloch, 1992):

$$\mathbf{Y}_T = \mathbf{Y}_B + \mathbf{Y}_W. \tag{14.3}$$

This decomposition can be used to compute a between-groups covariance matrix $\mathbf{\Sigma}_B$ (the population covariance matrix of the disaggregated group means $\mathbf{Y}_B$) and a within-groups covariance matrix $\mathbf{\Sigma}_W$ (the population covariance matrix of the individual deviations from the group means $\mathbf{Y}_W$). These covariance matrices are also orthogonal and additive:

$$\mathbf{\Sigma}_T = \mathbf{\Sigma}_B + \mathbf{\Sigma}_W. \tag{14.4}$$

Following the same logic, we can also decompose the sample data. Suppose we have data from $N$ individuals, divided into $G$ groups (subscript $i$ for individuals, $i = 1 \ldots N$; subscript $g$ for groups, $g = 1 \ldots G$). If we decompose the sample data, the sample covariance matrices are also orthogonal and additive:

$$\mathbf{S}_T = \mathbf{S}_B + \mathbf{S}_W. \tag{14.5}$$

Multilevel structural equation modeling assumes that the population covariance matrices $\mathbf{\Sigma}_B$ and $\mathbf{\Sigma}_W$ are described by distinct models for the between-groups and within-groups structure. To estimate the model parameters, the factor loadings, path coefficients, and residual variances, we need maximum likelihood estimates of the population between-groups covariance matrix $\mathbf{\Sigma}_B$ and the population within-groups covariance matrix $\mathbf{\Sigma}_W$. What we have is the observed between-groups matrix $\mathbf{S}_B$ and the observed within-groups matrix $\mathbf{S}_W$. It would be nice if we could simply construct a within-groups model for $\mathbf{S}_W$, and a between-groups model for $\mathbf{S}_B$. Unfortunately, we cannot simply use $\mathbf{S}_B$ as an estimate of $\mathbf{\Sigma}_B$, and $\mathbf{S}_W$ for $\mathbf{\Sigma}_W$. The situation is more complicated. Several different approaches have been offered for estimating multilevel factor models. This section describes three approaches: the pseudobalanced (MUML) approach, the two-phase direct estimation approach, and a weighted least squares approach.

### 14.1.2 Muthén's pseudobalanced approach

In the special case of balanced groups, meaning that all groups are the same size, estimation of a multilevel structural equation model is straightforward (Muthén, 1989). If we have $G$ balanced groups, with $G$ equal group sizes $n$ and a total sample size $N = nG$, we define two sample covariance matrices: the pooled within-covariance matrix $\mathbf{S}_{PW}$ and the scaled between-covariance matrix $\mathbf{S}_B^*$.

Muthén (1989) shows that an unbiased estimate of the population within-groups covariance matrix $\mathbf{\Sigma}_W$ is given by the pooled within-groups covariance matrix $\mathbf{S}_{PW}$, calculated in the sample by:

$$\mathbf{S}_{PW} = \frac{\sum_{j}^{G}\sum_{i}^{n}\left(\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_{j}\right)\left(\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_{j}\right)'}{N - G} \tag{14.6}$$

Equation 14.6 corresponds to the conventional equation for the covariance matrix of the individual deviation scores, with $N - G$ in the denominator instead of the usual $N - 1$, which reflects the loss of degrees of freedom by centering on the $G$ group means.

Since the pooled within-groups covariance matrix $\mathbf{S}_{PW}$ is an unbiased estimate of the population within-groups covariance matrix $\mathbf{\Sigma}_W$, we can estimate the population within-group model directly by constructing and estimating a model for $\mathbf{S}_{PW}$.

The scaled between-groups covariance matrix for the disaggregated group means $\mathbf{S}_B^*$, calculated in the sample, is given by:

$$\mathbf{S}_B^* = \frac{\sum_j^G n\left(\bar{\mathbf{Y}} - \bar{\mathbf{Y}}_j\right)\left(\bar{\mathbf{Y}} - \bar{\mathbf{Y}}_j\right)'}{G - 1}. \tag{14.7}$$

Muthén (1989, 1990) shows that $\mathbf{S}_{PW}$ is the maximum likelihood estimator of $\mathbf{\Sigma}_W$, with sample size $N - G$, and $\mathbf{S}_B^*$ is the maximum likelihood estimator of the composite $\mathbf{\Sigma}_W + c\mathbf{\Sigma}_B$, with sample size $G$, and $c$ equal to the common group size $n$:

$$\mathbf{S}_{PW} = \widehat{\mathbf{\Sigma}_W} \tag{14.8}$$

and:

$$\mathbf{S}_B^* = \widehat{\mathbf{\Sigma}_W + c\mathbf{\Sigma}_B}. \tag{14.9}$$

Equations 14.8 and 14.9 suggest using the multi-group option of conventional SEM software for a simultaneous analysis at both levels. However, if we model the between-groups structure, we cannot simply construct and test a simple model for $\mathbf{\Sigma}_B$, because $\mathbf{S}_B^*$ estimates a combination of $\mathbf{\Sigma}_W$ and $\mathbf{\Sigma}_B$. Instead, we have to specify for the $\mathbf{S}^*{}_B$ 'group' a model that contains two distinct sub-models: one for the within structure and one for the between structure. The procedure is that we specify two groups, with covariance matrices $\mathbf{S}_{PW}$ and $\mathbf{S}_B^*$ (based on $N - G$ and $G$ observations). The model for $\mathbf{\Sigma}_W$ must be specified for both $\mathbf{S}_{PW}$ and $\mathbf{S}_B^*$, with equality restrictions between both 'groups' to guarantee that we are indeed estimating the same model in both covariance matrices, and the model for $\mathbf{\Sigma}_B$ is specified for $\mathbf{S}_B^*$ only, with the scale factor $c$ built into the model.

The reasoning strictly applies only to the so-called *balanced* case, that is, if all groups have the same group size. In the balanced case, the scale factor $c$ is equal to the common group size $n$. The unbalanced case, where the group sizes differ, with $G$ groups of unequal sizes, is more complicated. In this case, $\mathbf{S}_{PW}$ is still the maximum likelihood estimator of $\mathbf{\Sigma}_W$, but $\mathbf{S}_B^*$ now estimates a different expression for each set of groups with distinct group size $d$:

$$\mathbf{S}^*_{Bd} = \widehat{\mathbf{\Sigma}_W + c_d\mathbf{\Sigma}_B}, \tag{14.10}$$

where equation 14.10 holds for each distinct set of groups with a common group size equal to $n_d$, and $c_d = n_d$ (Muthén, 1990, 1994). Full information maximum likelihood (FIML) estimation for unbalanced groups implies specifying a separate between-group model for each distinct group size. These between-groups models have different scaling parameters $c_d$ for each distinct group size, and require equality constraints across all other parameters and inclusion of a mean structure (Muthén, 1994, p. 385). Thus, using conventional SEM software for unbalanced data requires a complicated modeling scheme that creates a different 'group' for each set of groups with the same group size. This results in large and complex models, and possibly groups with a sample size less than the number of elements in the corresponding covariance matrix. This makes full maximum likelihood estimation problematic, and therefore Muthén (1989, 1990) proposes to ignore the unbalance, and to compute a single $\mathbf{S}^*_B$. The model for $\mathbf{S}^*_B$ includes an ad hoc estimator $c^*$ for the scaling parameter, which is close to the average sample size:

$$c^* = \frac{N^2 - \sum_j^G n_j^2}{N(G-1)}. \tag{14.11}$$

The result is a limited information maximum likelihood solution, which McDonald (1994) calls a pseudobalanced solution, and Muthén (1989, 1994) calls the MUML (for MUthén's ML) solution.

Figure 14.1 presents a path diagram for the pseudobalanced model. Left in the figure is the within model for the pooled within matrix. Right in the diagram is the model for the scaled between matrix. It repeats the model for the within matrix, with equality constraints across all corresponding parameters. On top of this, for each observed variable a corresponding latent between variable is specified. These between-level latent variables represent the second- or between-level intercept variances. $c$ is the scaling constant. The between-level model is constructed for the latent between variables.

Muthén (1989, 1990) shows that $\mathbf{S}^*_B$ is a consistent and unbiased estimator of the composite $\mathbf{\Sigma}_W + c\mathbf{\Sigma}_B$. This means that with large samples (of both individuals *and* groups!) $\mathbf{S}^*_B$ is a close estimate of $\mathbf{\Sigma}_B$, and the pseudobalanced solution produces a good approximation given adequate sample sizes.

Simulation studies (Hox and Maas, 2001; Hox, Maas, & Brinkhuis, 2010) find that the within-groups part of the model poses no problems in any of the simulated conditions. In the between-groups part of the model, the factor loadings are generally

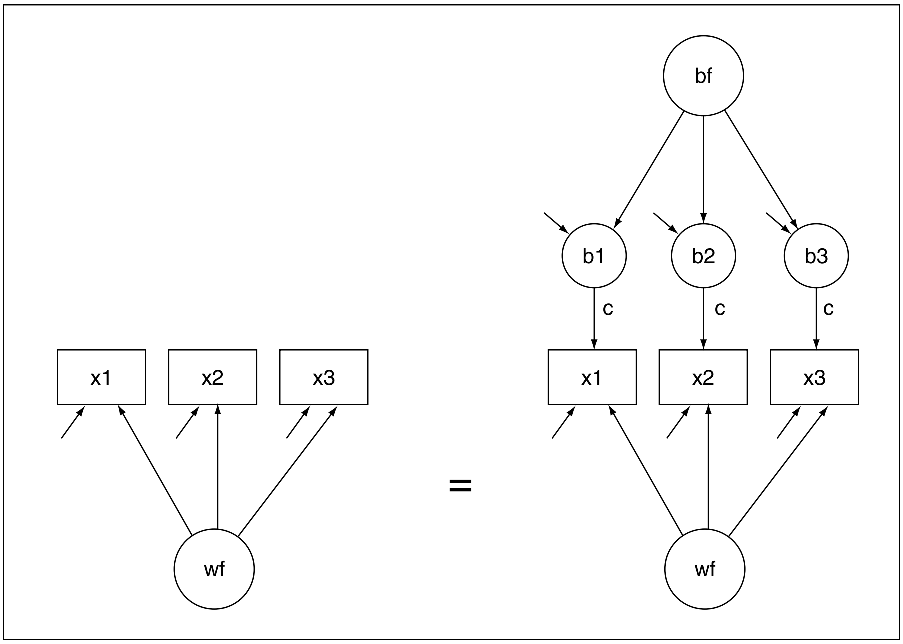


*Figure 14.1* Path diagram for pseudobalanced model.

accurate. However, the residual variances are underestimated, and the standard errors are generally too small, leading to an operating alpha level of about 8%. In addition, the chi-square model test is rejected too often, which again results in an operating alpha level of about 8%. Yuan and Hayashi (2005) show analytically that MUML standard errors and chi-square tests lead to correct inferences only when the between-level sample size goes to infinity and the coefficient of variation of the group sizes goes to zero. Both simulations and analytical work agree that larger sample sizes do not improve the accuracy with unbalanced data. So, with severely unbalanced data, MUML produces biased standard errors and significance tests, and this bias is not diminished when the sample size is increased.

Since the pseudobalanced approach needs the within-groups model for both the pooled within-groups model and the scaled between-groups model, and needs to incorporate the scaling factor for the between-groups model, the actual model can become quite complicated. Most modern structural equation software includes the MUML approach for two-level data, generally combined with default robust estimators for the standard errors and the chi-square test statistic to correct for the remaining heterogeneity (see Muthén & Satorra, 1995). The complications of the setup are generally hidden for the user. In software that does not include MUML as an option, it is still possible to calculate $\mathbf{S}_{PW}$ and $\mathbf{S}_B$ separately, and construct the complicated setup. For details I refer to Hox (1993) and the first edition of this book (Hox, 2002).

### 14.1.3 Direct estimation of the within and between matrix

Goldstein (1987, 2003) has suggested using the multivariate multilevel model described in Chapter 10 to produce a covariance matrix at the different levels, and to input these into a standard SEM program for further analysis. For our family data, we would use a multivariate multilevel model with three separate levels for the six intelligence tests, the individual children, and the families. We create six dummy variables to indicate the six intelligence scales, and exclude the intercept from the model. Hence, at the lowest level (the variables level) we have:

$$Y_{hij} = \pi_{1ij} d_{1ij} + \pi_{2ij} d_{2ij} + \ldots + \pi_{6ij} d_{6ij}, \tag{14.12}$$

at the individual level we have:

$$\pi_{pij} = \beta_{pj} + u_{pij}, \tag{14.13}$$

and at the family level (the third level in the multivariate model), we have:

$$\beta_{pj} = \gamma_p + u_{pj}. \tag{14.14}$$

By substitution we obtain:

$$\begin{aligned} Y_{hij} = {} & \gamma_1 d_{1ij} + \gamma_2 d_{2ij} + \ldots + \gamma_p d_{pij} \\ & + u_{1ij} d_{1ij} + u_{2ij} d_{2ij} + \ldots + u_{pij} d_{pij} + u_{1j} d_{1ij} + u_{2j} d_{2ij} + \ldots + u_{pj} d_{pij}. \end{aligned} \tag{14.15}$$

In sum notation, we have:

$$Y_{hij} = \sum_{h=1}^{6} \gamma_h d_{hij} + \sum_{h=1}^{6} u_{hij} d_{hij} + \sum_{h=1}^{6} u_{hj} d_{hij}. \tag{14.16}$$

The model described by equation 14.16 provides us with estimates of the six test means, and of their variances and covariances at the individual and family level in $\mathbf{\Sigma}_W$ and $\mathbf{\Sigma}_B$. Since in this application we are mostly interested in the variances and covariances, RML estimation is preferred to FML estimation. The individual-level (within) covariances and the family-level (between) covariances and means are direct maximum likelihood estimators of their population counterparts, which can be input to any SEM software, either separately or in a two-group analysis. Hox and Maas (2004) explain this approach in more detail.

The fact that the individual-level (within families) and family-level (between families) covariances are estimated directly, and consequently can be modeled directly and separately by any SEM program, is a distinct advantage of the multivariate multilevel approach. As a result, we routinely get separate model tests and fit indices at all levels. The multivariate multilevel approach to multilevel SEM also generalizes straightforwardly to more than two levels. There are other advantages as well. First, since the multilevel multivariate model does not assume that we have a complete set of variables for each individual, incomplete data are accommodated without special effort. Second, if we have dichotomous or ordinal variables, we can use the multilevel generalized linear model to produce the covariance matrices, again without special effort.

There are some disadvantages to the multivariate multilevel approach as well. An important disadvantage is that the covariances produced by the multivariate multilevel approach are themselves estimated values. They are not directly calculated, like the pooled within-groups and scaled between-groups covariances are, but they are estimates produced by a complex statistical procedure. The estimation in the second step treats these estimates as data or observed covariance matrices. If the data have a multivariate normal distribution, the within-groups and between-groups covariances can be viewed as observed values, which have a known sampling distribution. However, when we have incomplete data it is unclear what the proper sample size is, and in the case of non-normal (e.g., dichotomous or ordinal) variables we know that the sampling distribution is not normal. Since the normal sampling distribution is used by SEM programs to calculate the chi-square model test and the standard errors of the parameter estimates, the chi-square test and standard errors cannot be trusted unless data are complete and multivariate normal.

### 14.1.4 Analysis of the within and between matrix using weighted least squares

Asparouhov and Muthén (2007) describe an approach to multilevel SEM that uses separate estimation of the covariance matrices followed by an estimation method that overcomes the problems that are encountered with direct estimation of the within and between covariance matrix with non-normal data. In this approach, univariate maximum likelihood methods are used to estimate the vector of means $\boldsymbol{\mu}$ at the between-group level, and the diagonal elements (the variances) of $\boldsymbol{\Sigma}_W$ and $\boldsymbol{\Sigma}_B$. In the case of ordinal categorical variables, thresholds are estimated as well. Next, the off-diagonal elements of $\boldsymbol{\Sigma}_W$ and $\boldsymbol{\Sigma}_B$ are estimated using bivariate maximum likelihood methods. Finally, the asymptotic variance–covariance matrix for these estimates is obtained, and the multilevel SEM is estimated for both levels using weighted least squares (WLS). Currently, this approach is only available in Mplus.

WLS is an estimation method that uses the sampling variance–covariance matrix of $\mathbf{S}_W$ and $\mathbf{S}_B$ as a weight matrix to obtain correct chi-squares and standard errors. This estimation method is developed for efficient estimation of multilevel models with non-normal variables, since for such data full maximum likelihood estimation requires high-dimensional numerical integration, which is computationally very demanding. Multilevel WLS can also be used for multilevel estimation with continuous variables, but then it does not have a real advantage.

Standard WLS uses a weight matrix based on the asymptotic covariances of all estimated parameters. For the unrestricted model, the number of parameters is large, and the asymptotic covariance matrix is very large. Especially for the between part of the model, the number of elements in this matrix can easily become larger than the number of groups. Unless the number of groups is very large, it is preferable to use only the diagonal of this matrix, (see Muthén, Du Toit, & Spisic, 1997), an approach known as diagonal WLS (DWLS). In Mplus, choosing the diagonal weight matrix always implies using a robust chi-square (WLSM using a mean-corrected (first order) or WLSMV using a mean-and-variance corrected (second order) correction).

## 14.2 FULL MAXIMUM LIKELIHOOD ESTIMATION

In two-level data, the factor structure given by 14.1 becomes:

$$\begin{aligned} \mathbf{y}_W &= \boldsymbol{\Lambda}_W \boldsymbol{\eta}_W + \boldsymbol{\varepsilon}_W \\ \boldsymbol{\mu}_B &= \boldsymbol{\mu}_B + \boldsymbol{\Lambda}_B \boldsymbol{\eta}_B + \boldsymbol{\varepsilon}_B \end{aligned} \tag{14.17}$$

where $\boldsymbol{\mu}_B$ are the random intercepts that vary across groups. The first equation models the within-groups variation, and the second equation models the between-groups variation. By substitution and rearranging terms we obtain:

$$\mathbf{y}_{ij} = \boldsymbol{\mu}_B + \boldsymbol{\Lambda}_W \boldsymbol{\eta}_W + \boldsymbol{\Lambda}_B \boldsymbol{\eta}_B + \boldsymbol{\varepsilon}_B + \boldsymbol{\varepsilon}_W. \tag{14.18}$$

In 14.18, $\boldsymbol{\mu}$ is the vector of means, $\boldsymbol{\Lambda}_W$ is the factor matrix at the within level, $\boldsymbol{\Lambda}_B$ is the factor matrix at the between level, and $\boldsymbol{\varepsilon}_W$ and $\boldsymbol{\varepsilon}_B$ are the residual errors at the within and the between level. Except for the notation, the structure of equation 14.18 follows that of a random intercept regression model, with fixed regression coefficients in the factor matrices $\boldsymbol{\Lambda}$ and a level 1 and level 2 error term. By allowing group-level variation in the factor loadings we can generalize this to a random coefficient model. In the context of multilevel factor analysis, varying loadings are problematic because they imply that the measurement model is not equivalent across the different groups. In the context of multilevel path analysis, random coefficients for relationships between variables provide information on differences between groups that have a substantive interpretation.

To provide maximum likelihood estimates for the parameters in 14.18 in the general case of unbalanced groups, we need to model the observed raw data. The two-stage approaches all follow the conventional notion that structural equation models are constructed for the covariance matrix with possibly an added mean vector. When data have a multivariate normal distribution, these are the sufficient statistics, and raw data are superfluous. Thus, for a confirmatory factor model, the covariance matrix $\mathbf{\Sigma}$ is:

$$\mathbf{\Sigma} = \mathbf{\Lambda\Phi\Lambda}' + \mathbf{\Theta}, \tag{14.19}$$

where $\mathbf{\Lambda}$ is the factor matrix, $\mathbf{\Phi}$ is the covariance matrix of the latent variables, and $\mathbf{\Theta}$ is the vector with residual variances. The parameters are commonly estimated by maximizing the likelihood function, or equivalently minimizing the discrepancy function (Bollen, 1989):

$$F = \log|\mathbf{S}| + \mathrm{tr}(\mathbf{SS}^{-1}) - \log|\mathbf{S}| - p, \tag{14.20}$$

where '|.|' indicates the determinant of a matrix, 'tr' indicates the trace, and $p$ is the number of manifest variables.

Unbalanced groups can be viewed as a form of incomplete data. For incomplete data, the maximum likelihood approach defines the model and the likelihood in terms of the raw data, which is why it is sometimes called the raw likelihood method. Raw ML minimizes the function (Arbuckle, 1996):

$$F = \sum_{i=1}^{N} \log|\mathbf{S}_i| + \sum_{i=1}^{N} \log(\mathbf{x}_i - \boldsymbol{\mu}_i)'\mathbf{S}_i^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_i), \tag{14.21}$$

where the subscript $i$ refers to the observed cases, $\mathbf{x}_i$ to those variables observed for case $i$, and $\boldsymbol{\mu}_i$ and $\mathbf{\Sigma}_i$ contain the population means and covariances of the variables observed for case $i$. If the data are complete, equations 14.20 and 14.21 are equivalent. If the data are incomplete, the covariance matrix is no longer a sufficient statistic, and minimizing the discrepancy function given by 14.21 provides the maximum likelihood estimates for the incomplete data.

Mehta and Neale (2005) show that models for multilevel data, with individuals nested within groups, can be expressed as a structural equation model. The fit function given by equation 14.21 applies, with clusters as units of observation, and individuals within clusters as variables. Unbalanced data, here unequal numbers of individuals within clusters, are included in the same way as incomplete data in standard SEM. While the two-stage approaches (MUML, direct estimation, and WLS) include only random intercepts in the between-groups model, the full ML representation can incorporate random slopes as well (Mehta & Neale, 2005). In theory, any modern SEM software that allows for incomplete data can be used to estimate multilevel SEM. In practice, specialized software is used that makes use of the specific multilevel structure in the data to simplify calculations. Full maximum likelihood multilevel SEM is currently available for two-level models in Mplus and for multilevel models in GLLAMM. A recent development is to use robust standard errors and chi-squares for significance testing. With multilevel data, robust chi-squares and standard errors offer some protection against unmodeled heterogeneity, which may result from misspecifying the group level model, or by omitting a level. Finally, Skrondal and Rabe-Hesketh (2004) show how to combine this with a generalized linear model for the observed variables, which allows for non-normal variables. This is currently available only in Mplus and GLLAMM.

The next section analyzes an example data set, employing all four estimation methods discussed above. When all methods are compared, the pseudobalanced MUML approach and the multivariate multilevel two-stage estimation method are the least accurate, especially for the between-model estimates. These methods are clearly outdated. Simulations (Hox et al., 2010) have shown that both WLS and full ML estimation are more accurate than MUML, and that the difference between WLS and ML is negligible. Our example confirms that ML and WLS are quite similar. When ML estimation is possible, it is the method of choice, but when the demands for ML estimation overtax the computer capacity, WLS is a viable alternative.

The maximum likelihood approach is the only approach that allows random slopes in the model. In a confirmatory factor analysis, this means that factor loadings are allowed to vary across groups. In our example, none of the six individual-level factor loadings has significant slope variation across families. In confirmatory factor analysis this is desirable, because finding varying factor loadings implies that the scales involved are not measuring the same thing in the same way across families. Varying slopes in a measurement model indicate a lack of measurement equivalence across groups.

## 14.3 AN EXAMPLE OF MULTILEVEL FACTOR ANALYSIS

The example data are the scores on six intelligence measures of 400 children from 60 families, patterned after Van Peet (1992). The six intelligence measures are: word list, cards, matrices, figures, animals, and occupations. The data have a multilevel structure, with children nested within families. If intelligence is strongly influenced by shared genetic and environmental influences in the families, we may expect rather strong between-family effects. In this data set, the intraclass correlations of the intelligence measures range from .38 to .51, which are indeed large intraclass correlations.

### 14.3.1 Full maximum likelihood estimation

Given that full maximum likelihood estimation is the norm, we begin the analysis of the example data using this estimation method. Muthén (1994) recommends starting the analysis with an analysis of the total scores. This may have been good advice when the complicated pseudobalanced model setups were used, but given user-friendly multilevel SEM software, this step is superfluous. Since the effective level 1 sample size ($N - G$) is almost always much larger than the level 2 sample size ($G$), it is good practice to start with the within part, either by specifying a saturated model for the between part, or by analyzing only the pooled within matrix.

In the example data, the number of observations on the individual level are 400 – 60 = 340, while on the family level there are 60. Thus, it makes sense to start at the individual level by constructing a model for $\mathbf{S}_{PW}$ only, ignoring $\mathbf{S}_B$.

An exploratory factor analysis on the correlations derived from $\mathbf{S}_{PW}$ suggests two factors, with the first three measures loading on the first factor, and the last three measures on the last. A confirmatory factor analysis on $\mathbf{S}_{PW}$ confirms this model ($\chi^2 =$ 6.0, $df = 8$, $p = .56$). A model with just one general factor for $\mathbf{S}_{PW}$ is rejected ($\chi^2 = 207.6$, $df = 9$, $p < .001$).

The next step is to specify a family model. For explorative purposes, we could carry out a separate analysis on the estimated between-groups covariance matrix $\mathbf{S}_B$. This matrix, which is a maximum likelihood estimator of $\mathbf{\Sigma}_B$ (not the scaled between-matrix produced by 14.7), is obtained by specifying a saturated model for both the within and the between level (Mplus generates these matrices automatically). In the example data, given the good fit of the within model, we carry on with an analysis of the multilevel data, with the two-factor model retained for the within part.

We start the analysis of the between structure by estimating some benchmark models for the group level, to test whether there is any between family structure at all. The simplest model is a null model that completely leaves out the specification of a family-level model. If the null model holds, there is no family-level structure at all; all covariances in $\mathbf{S}_B$ are the result of individual sampling variation. If this null model holds, we may as well continue our analyses using simple single-level analysis methods.

The next model tested is the independence model, which specifies only variances on the family level, but no covariances. The independence model estimates for the family-level structure only the variances of the family-level variables. If the independence model holds, there is family-level variance, but no substantively interesting structural model. We can simply analyze the pooled within matrix, at the cost of losing the within-group information from *G* observations that is contained in the between-groups covariance matrix. If the independence model is rejected, there is some kind of structure at the family level. To examine the best possible fit given the individual-level model, we can estimate the saturated model that fits a full covariance matrix to the family-level observations. This places no restrictions on the family model. Table 14.1 shows the results of estimating these benchmark models on the family level.

*Table 14.1* Family-level benchmark models

| Family model | Chi-square | *df* | *p* |
|---|---|---|---|
| Null | 323.6 | 29 | .00 |
| Independence | 177.2 | 23 | .00 |
| Saturated | 6.7 | 8 | .57 |

The null model and the independence model are both rejected. Subsequently, we specify for the family level the same one-factor and two-factor models we have used at the individual level. The one-factor model fits well ($\chi^2$ = 11.9, *df* = 17, *p* = .80). The two-factor model is no improvement (difference chi-square 0.15, *p* = .70).

The principle of using the simplest model that fits well leads to acceptance of the one-factor model at the family level. The chi-square model test is not significant, and the fit indices are fine: CFI is 1.00 and the RMSEA is .00. Figure 14.2 presents the within and between model in a single path diagram. Note that, consistent with the full path diagram in Figure 14.1, the between-level variables that represent the family-level intercept variance of the observed variables are latent variables, depicted by circles or ellipses. Using the maximum likelihood method (ML) in Mplus leads to the estimates reported in Table 14.2.

### 14.3.2 Weighted least squares estimation

Using the separate estimation/WLS method with robust chi-square (WLSMV) in Mplus leads to the estimates reported in Table 14.3.[1] The chi-square model test accepts the model ($\chi^2$ = 5.91, *df* = 7, *p* = .55), the fit indices are good: CFI is 1.00 and the

[1] The asymptotic chi-square is available by specifying full WLS estimation. As explained earlier, this leads to a very large weight matrix. With a group-level sample size of only 60 this is a recipe for disaster, hence the choice for robust WLS estimation.

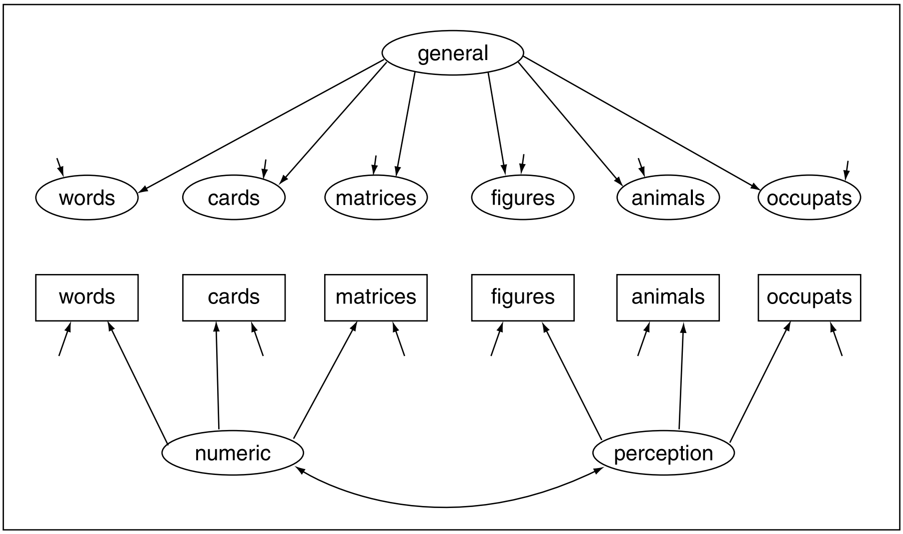


*Figure 14.2* Diagram for family IQ data 1.

*Table 14.2* Individual- and family-level estimates, ML estimation

| | Individual level | | | Family level | |
|---|---|---|---|---|---|
| | Numer. | Percept. | Resid. var. | General | Resid. var. |
| Wordlst | 3.18 (.20) | | 6.19 (.74) | 3.06 (.39) | 1.25 (.57) |
| Cards | 3.14 (.19) | | 5.40 (.69) | 3.05 (.39) | 1.32 (.59) |
| Matrix | 3.05 (.20) | | 6.42 (.71) | 2.63 (.38) | 1.94 (.67) |
| Figures | | 3.10 (.20) | 6.85 (.76) | 2.81 (.40) | 2.16 (.71) |
| Animals | | 3.19 (.19) | 4.88 (.70) | 3.20 (.38) | 0.66 (.49) |
| Occupat | | 2.78 (.18) | 5.33 (.60) | 3.44 (.42) | 1.58 (.62) |

Standard errors in parentheses. Correlation between individual factors: .38.

RMSEA is .00. The parameter estimates in Table 14.3 are similar to the full maximum likelihood estimates, but not identical. The robust standard errors lead to the same conclusions as the asymptotic standard errors used with full maximum likelihood estimation.

*Table 14.3* Individual- and family-level estimates, WLS estimation

| | Individual level | | | Family level | |
|---|---|---|---|---|---|
| | Numer. | Percept. | Resid. var. | General | Resid. var. |
| Wordlst | 3.25 (.15) | | 5.67 (.84) | 3.01 (.48) | 1.51 (.62) |
| Cards | 3.14 (.18) | | 5.44 (.68) | 3.03 (.38) | 1.25 (.71) |
| Matrix | 2.96 (.22) | | 6.91 (.92) | 2.62 (.45) | 2.02 (.69) |
| Figures | | 2.96 (.22) | 7.67 (.92) | 2.80 (.46) | 2.03 (.72) |
| Animals | | 3.35 (.21) | 3.79 (.99) | 3.15 (.41) | 0.96 (.61) |
| Occupat | | 2.75 (.24) | 5.49 (.94) | 3.43 (.44) | 1.67 (.63) |

Standard errors in parentheses. Correlation between individual factors: .38.

### 14.3.3 Direct estimation of the within and between matrix followed by separate modeling

Since the family example data are multivariate normal and complete, the direct estimation method discussed in section 14.2.3 is a viable option. Covariance matrices were estimated for the within and the between part using the multivariate multilevel model presented in Chapter 10. Estimation was carried out in MLwiN restricted maximum likelihood; subsequent SEM modeling was done in AMOS. The within model does fit well: $\chi^2 = 6.9$, $df = 8$, $p = .55$, CFI is 1.00 and the RMSEA is .00. Analysis of the between matrix produces inconsistent results: $\chi^2 = 15.8$, $df = 9$, $p < .07$, CFI is .99 but the RMSEA is .11. The parameter estimates for the model are presented in Table 14.4.

*Table 14.4* Individual- and family-level estimates, direct estimation

| | Individual level | | | Family level | |
|---|---|---|---|---|---|
| | Numer. | Percept. | Resid. var. | General | Resid. var. |
| Wordlst | 3.14 (.20) | | 6.21 (.73) | 3.04 (.32) | 1.52 (.33) |
| Cards | 3.15 (.20) | | 5.32 (.69) | 3.07 (.32) | 1.13 (.26) |
| Matrix | 3.03 (.20) | | 6.40 (.71) | 2.64 (.31) | 1.16 (.43) |
| Figures | | 3.11 (.21) | 6.80 (.75) | 2.81 (.32) | 2.03 (.41) |
| Animals | | 3.20 (.19) | 4.84 (.68) | 3.18 (.32) | 0.82 (.22) |
| Occupat | | 2.81 (.18) | 5.34 (.60) | 3.44 (.36) | 1.64 (.39) |

Standard errors in parentheses. Correlation between individual factors: .38.

Especially for the between part and its standard errors, the parameter estimates obtained by direct estimation differ to some extent from the ML and WLS estimates. Hox and Maas (2004) find better results for direct estimation, but they use a very large sample and a model with small error variances. In this example, it appears that direct estimation of $\mathbf{\Sigma_W}$ and $\mathbf{\Sigma_B}$ is less accurate than the full ML or WLS method.

### 14.3.4 Pseudobalanced (MUML) estimation

Using the pseudobalanced (MUML) estimation in Mplus leads to the estimates reported in Table 14.5. The chi-square model test accepts the model ($\chi^2 = 11.3$, $df = 17$, $p = .8455$) and the fit indices are good: CFI is 1.00 and the RMSEA is .00. The parameter estimates in Table 14.5 are similar to the full maximum likelihood estimates, but not identical. The standard errors lead to the same conclusions as the asymptotic standard errors used with full maximum likelihood estimation.

*Table 14.5* Individual- and family-level estimates, MUML estimation

| | Individual level | | | Family level | |
|---|---|---|---|---|---|
| | Numer. | Percept. | Resid. var. | General | Resid. var. |
| Wordlst | 3.16 (.20) | | 6.24 (.74) | 2.95 (.38) | 1.32 (.58) |
| Cards | 3.15 (.19) | | 5.36 (.69) | 2.97 (.38) | 1.35 (.57) |
| Matrix | 3.06 (.20) | | 6.37 (.72) | 2.60 (.37) | 1.95 (.66) |
| Figures | | 3.10 (.20) | 6.85 (.76) | 2.77 (.39) | 2.19 (.71) |
| Animals | | 3.19 (.19) | 4.81 (.68) | 3.12 (.37) | 0.70 (.46) |
| Occupat | | 2.78 (.18) | 5.33 (.60) | 3.31 (.40) | 1.59 (.61) |

Standard errors in parentheses. Correlation between individual factors: .38.

The MUML method produces estimates and standard errors that are again somewhat different from the full ML and WLS estimates, especially for the between part. Given that better methods exist, it should be considered outdated and superseded by full ML and WLS estimation.

## 14.4 STANDARDIZING ESTIMATES IN MULTILEVEL STRUCTURAL EQUATION MODELING

The estimates reported earlier are all unstandardized estimates. For interpretation, it is often useful to inspect the standardized estimates as well, because these can be used to compare the loadings and residual variances for variables that are measured in a different metric. A convenient standardization is to standardize both the latent factors and the observed variables on each level separately. Table 14.6 presents the standardized estimates for the ML estimates. It shows that the factor structure at the family level is stronger than at the individual level. This is typical; one reason is that measurement errors accumulate at the individual level.

*Table 14.6* Individual- and family-level estimates, standardized estimates

| | Individual level | | | Family level | |
|---|---|---|---|---|---|
| | Numer. | Percept. | Resid. var. | General | Resid. var. |
| Wordlst | 0.79 (.03) | | 0.38 (.05) | 0.94 (.03) | 0.12 (.06) |
| Cards | 0.80 (.03) | | 0.35 (.05) | 0.94 (.03) | 0.12 (.06) |
| Matrix | 0.77 (.03) | | 0.41 (.05) | 0.88 (.05) | 0.22 (.08) |
| Figures | | 0.76 (.03) | 0.42 (.05) | 0.89 (.04) | 0.22 (.08) |
| Animals | | 0.82 (.03) | 0.32 (.05) | 0.97 (.02) | 0.06 (.05) |
| Occupat | | 0.77 (.03) | 0.40 (.05) | 0.94 (.03) | 0.12 (.05) |

Standard errors in parentheses. Correlation between individual factors: .38.

The separate standardization presented in Table 14.6 is called the within-groups completely standardized solution. Standardization takes place separately in the within part and in the between part. Some software also produces a ‘common metric solution’, which standardizes the variables using a common metric across the groups or levels. In multilevel SEM, this solution produces standardized estimates that have no meaningful interpretation. Gustafsson and Stahl (1999) propose a different standardization, built into their preprocessor STREAMS. This is a multilevel standardization that standardizes the latent variables at each separate level, but uses the total standard deviation of the observed variables to standardize both the within-groups and the between-groups level. This provides a better insight into how much variance each factor explains at the different levels. This multilevel standardization can be accomplished by hand calculation using other SEM software.[2] Summarizing: Some careful thought is needed in choosing the correct standardization method.

## 14.5 GOODNESS OF FIT IN MULTILEVEL STRUCTURAL EQUATION MODELING

SEM programs produce in addition to the chi-square test a number of goodness-of-fit indices that indicate how well the model fits the data. Statistical tests for model fit have the problem that their power varies with the sample size. If we have a very large sample, the statistical test will almost certainly be significant. Thus, with large samples, we will always reject our model, even if the model actually describes the data quite well. Conversely, with a very small sample, the model will always be accepted, even if it fits rather badly.

Given the sensitivity of the chi-square statistic to the sample size, researchers have proposed a variety of alternative fit indices to assess model fit. All goodness-of-fit measures are some function of the chi-square and the degrees of freedom. Most of these fit indices do not only consider the fit of the model, but also its simplicity. A saturated model that specifies all possible paths between all variables always fits the data perfectly, but it is just as complex as the observed data. In general, there is a trade-off between the fit of a model and the simplicity of a model. Several goodness-of-fit indices assess simultaneously both the fit and the simplicity of a model. The goal is to produce a goodness-of-fit index that does not depend on the sample size or the distribution of the data. In fact, simulations have shown that most goodness-of-fit indices still depend on sample size and distribution, but the dependency is much smaller than that of the routine chi-square test.

Most SEM software computes a bewildering array of goodness-of-fit indices. All of them are functions of the chi-square statistic, but some include a second function that penalizes complex models. For instance, Akaike's information criterion (AIC) is twice the chi-square statistic minus twice the degrees of freedom for the model. For a detailed review and evaluation of a large number of fit indices, including those mentioned here, I refer to Gerbing and Anderson (1992).

Jöreskog and Sörbom (1989) have introduced two goodness-of-fit indices called GFI (goodness of fit) and AGFI (adjusted GFI). The GFI indicates goodness of fit, and the AGFI attempts to adjust the GFI for the complexity of the model. Bentler (1990) has introduced a similar index called the comparative fit index CFI. Two other well-known fit measures are the Tucker-Lewis index (TLI) (Tucker & Lewis, 1973), also

[2] Loadings are standardized by dividing by the standard deviation of the variables, variances by dividing by the square of the standard deviation.

known as the non-normed fit index or NNFI, and the normed fit index (NFI) (Bentler & Bonett, 1980). Both the NNFI and the NFI adjust for complexity of the model. Simulation research shows that all these indices still depend on sample size and estimation method (e.g., ML or GLS), with the CFI and the TLI/NNFI showing the best overall performance (Chou & Bentler, 1995; Kaplan, 1995). If the model fits perfectly, these fit indices should have the value 1. Usually, a value of at least .90 is required to accept a model, while a value of at least .95 is required to judge the model fit as 'good'. However, these are just rules of thumb.

A different approach to model fit is to accept that models are only approximations, and that a perfect fit may be too much to ask for. Instead, the problem is to assess how well a given model approximates the true model. This view led to the development of an index called RMSEA, for root mean square error of approximation (Browne & Cudeck, 1992). If the approximation is good, the RMSEA should be small. Typically, an RMSEA of less than .08 is required (Kline, 2004), and an RMSEA of less than .05 is required to judge the model fit as 'good'. Statistical tests or confidence intervals can be computed to test if the RMSEA is significantly larger than this lower bound.

Given the many possible goodness-of-fit indices, the customary advice is to assess fit by inspecting several fit indices that derive from different principles. Therefore, for the confirmatory factor model for the family data, I have reported the chi-square test, and the fit indices CFI and RMSEA.

A general problem with these goodness-of-fit indices in multilevel SEM is that they apply to the entire model. Therefore, the goodness-of-fit indices reflect the degree of fit both in the within model and in the between model. Since the sample size for the within 'group' is generally the largest, this part of the model dominates the value of the fit indices. Clearly, it makes sense to assess the fit for both parts of the model separately.

Since the within-groups sample size is usually much larger than the between-groups sample size, we do not lose much information if we model the within-groups matrix separately, and interpret the fit indices produced in this analysis separately.

A simple way to obtain goodness-of-fit indices for the between model is to specify a saturated model for the within-groups level. The saturated model estimates all covariances between all variables. It has no degrees of freedom, and always fits the data perfectly. As a result, the degree of fit indicated by the goodness-of-fit indices represents the (lack of) fit of the between model. This is not the best way to assess the fit of the between model, because the perfect fit of the within model also influences the value of the fit index. Fit indices that are mostly sensitive to the degree of fit will show a spuriously good fit, while fit indices that also reflect the parsimony of the model may show a spurious lack of fit.

A better way to indicate the fit of the within and between model separately is to calculate these by hand. Most fit indices are a simple function of the chi-square, sample size *N*, and degrees of freedom *df*. Some consider only the current model, the target model $M_t$; others also consider a baseline model, usually the independence model $M_I$. By estimating the independence and the target model for the within matrix, with a saturated model for the between matrix, we can assess how large the contribution to the overall chi-square is for the various within models. In the same way, by estimating the independence and the target model for the between matrix, with a saturated model for the within matrix, we can assess how large the contribution to the overall chi-square is for the various between models. Using this information, we can calculate the most common goodness-of-fit indices. Most SEM software produces the required information, and the references and formulas are in the user manuals and in the general literature (e.g., Gerbing & Anderson, 1992).

Table 14.7 gives the separate chi-squares, degrees of freedom, and sample sizes for the independence model and the final model for the family intelligence example.

*Table 14.7* Chi-squares and *df* for individual- and family-level models separately

| | Individual level, between model saturated | | Family level, within model saturated | |
|---|---|---|---|---|
| | Independence | 2 factors | Independence | 1 factor |
| Chi-square | 805.51 | 6.72 | 168.88 | 4.74 |
| *df* | 30 | 8 | 15 | 9 |
| *N* | 340 | 340 | 60 | 60 |

The comparative fit index CFI (Bentler, 1990) is given by:

$$CFI = 1 - \frac{\chi_t^2 - df_t}{\chi_I^2 - df_I}. \tag{14.22}$$

In equation 14.22, $\chi_t^2$ is the chi-square of the target model, $\chi_I^2$ is the chi-square for the independence model, and $df_t$ and $df_I$ are the degrees of freedom for the target and the independence model. If the difference between the chi-square and the degrees of freedom is negative, it is replaced by zero. So, for example, the CFI for the family-level model is given by:

$$CFI = 1 - (4.74 - 9)/(168.88 - 15) = 1 - 0/153.88 = 1.00.$$

The Tucker-Lewis index, TLI, which is also known as the non-normed fit index, NNFI, is given by:

$$TLI = \frac{\frac{\chi_I^2}{df_I} - \frac{\chi_t^2}{df_t}}{\frac{\chi_I^2}{df_I} - 1}. \tag{14.23}$$

Finally, the root mean square error of approximation, RMSEA, is given by:

$$RMSEA = \sqrt{\left(\frac{\chi_t^2 - df_t}{Ndf_t}\right)} \tag{14.24}$$

where $N$ is the total sample size. If RMSEA is negative, it is replaced by zero. Using equations 14.22 to 14.24 and the values in Table 14.7, we can calculate the CFI, TLI, and RMSEA separately for the within and between models. The results are shown in Table 14.8.

*Table 14.8* Fit indices for individual- and family-level models separately

| | Individual level, 2 factors | Family level, 1 factor |
|---|---|---|
| CFI | 1.00 | 1.00 |
| TLI | 1.01 | 1.05 |
| RMSEA | .00 | .00 |

The goodness-of-fit indices in Table 14.8 all indicate excellent fit, for both the within and the between model.

## 14.6 NOTATION AND SOFTWARE

Most modern SEM software includes routines for two-level SEM. Having only two levels may seem an important limitation, but one must appreciate that SEM is an inherently multivariate technique, and multilevel regression is univariate. Consequently, for multivariate analysis or including a measurement model, multilevel regression needs an extra ‘variable’ level, and for longitudinal analysis it needs an ‘occasion’ level. Multilevel SEM does not need this.

Nevertheless, having only two levels can put strong limits on the analysis. At the time, only the software GLLAMM (Rabe-Hesketh & Skrondal, 2008) can analyze multiple-level SEM. Using direct estimation as described above, more than two levels can be accommodated, but this is restricted to multivariate normal variables, and the example shows that the estimates and standard errors are not very accurate.

The two-stage approaches are simpler than the general random coefficient model. They are comparable to the multilevel regression model with random variation only for the intercepts. There is no provision for randomly varying slopes (factor loadings and path coefficients). Although it would be possible to include cross-level interactions, introducing interaction variables of any kind in structural equation models is complicated (see Bollen, 1989; Marcoulides & Schumacker, 2001). An interesting approach is allowing different within-groups covariance matrices in different sub-samples, by combining two-level and multigroup models.

When maximum likelihood estimation is used, multilevel SEM can include varying slopes. At the time, only Mplus and GLLAMM support this. Muthén and Muthén (1998–2007) have extended the standard path diagram by using a black dot on an arrow in the level 1 model to indicate a random slope. This slope appears in the level 2 model as a latent variable. This is consistent with the use of latent variables for the level 2 intercept variances. This highlights an important link between multilevel regression and multilevel SEM: random coefficients are latent variables, and many multilevel regression models can also be specified in the SEM context (Curran, 2003; Mehta & Neale, 2005).

Figure 14.3 shows an example of a path diagram from the Mplus manual (Muthén & Muthén, 1998–2007, p. 232). The within model depicts a simple regression

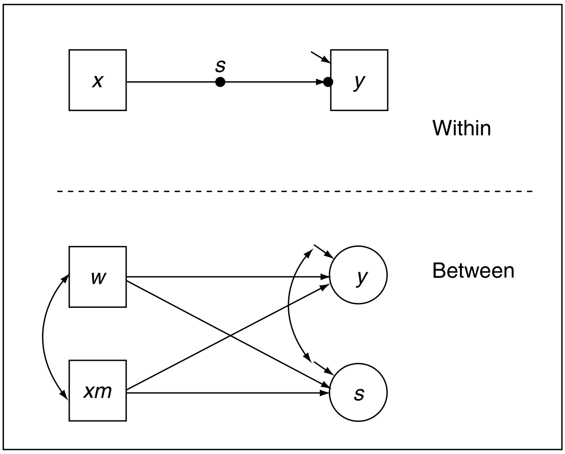


*Figure 14.3* Example of path model with random slope and intercept.

of the outcome variable *y* on the predictor variable *x*. The black dot on *y* indicates a random intercept for *y*, which is referred to as *y* in the between part of the model. The black dot on the arrow from *x* to *y* indicates a random slope, which is referred to as *s* in the between part of the model. In the between part of the model, there are two predictors that are measured only at the between level: the group-level variable *w* and the group mean on the variable *x* which is referred to as *xm*. Curran and Bauer (2007) extend this 'black dot' notation by placing a circle on arrows that indicate a random coefficient, where a circle is the general symbol for a latent variable. These latent variables can then be used in the path diagram like any other latent variable. This nicely underlines that a random coefficient is a latent variable.

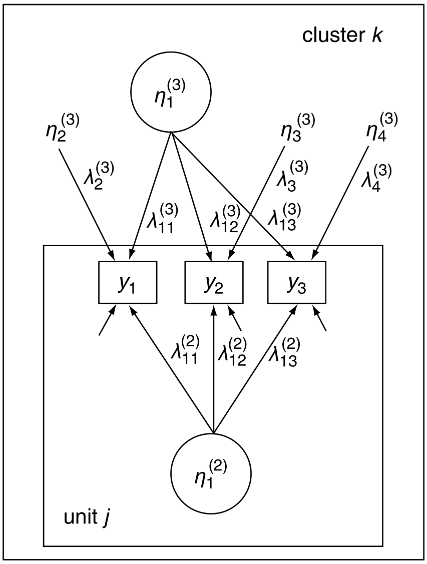


*Figure 14.4* Example of multilevel factor model.

Figure 14.4 is an example of a multilevel path diagram from Skrondal and Rabe-Hesketh (2004, p. 104). This is a three-level model, with items at the lowest level. The path diagram shows the unit (2nd) and cluster (3rd) level. This notation is a little more complicated, but is more easily extended to more complicated models, such as a partial nesting structure.

# 15

# Multilevel Path Models

Path models are structural equation models that consist of complex paths between latent and/or observed variables, possibly including both direct and indirect effects, and reciprocal effects between variables. As mentioned in Chapter 14, often a distinction is made between the structural and the measurement part of a model. The measurement model specifies how the latent factors are measured by the observed variables, and the structural model specifies the structure of the relationships between the theoretical constructs, which may be latent factors or observed variables in the model.

A multilevel path model uses the same approaches outlined in Chapter 14 for multilevel factor analysis. Chapter 14 discusses several different estimation methods, but in this chapter maximum likelihood is used throughout.

With multilevel path models, we often have the complication that there are pure group-level variables (*global* variables in the terminology of Chapter 1). An example would be the global variable *group size*. This variable simply does not exist at the individual level. We can of course disaggregate group size to the individual level. However, this disaggregated variable is constant within each group, and as a result the variance and covariances with the individual deviation scores are all zero. Actually, what we have in this case is a different set of variables at the individual and the group level. Some SEM software can deal directly with groups or levels that do not have the same variables. Estimation is not a problem with such software. Some SEM software (e.g., LISREL) requires that both groups or levels have the same variables. This problem can be solved by viewing the group-level variable as a variable that is systematically missing in the individual-level data. Bollen (1989) and Jöreskog and Sörbom (1989) describe how systematically missing variables can be handled in LISREL. The details are discussed in the appendix to this chapter.

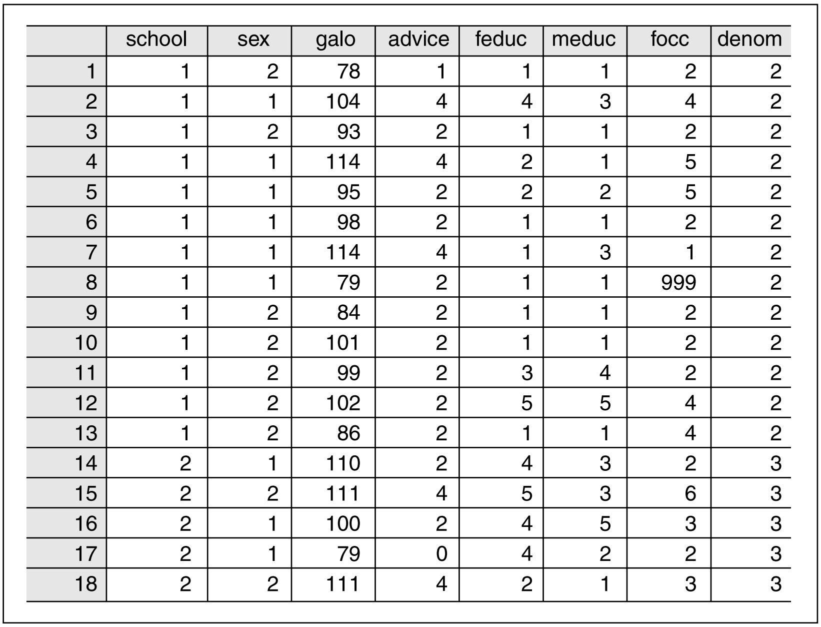
## 15.1 EXAMPLE OF A MULTILEVEL PATH ANALYSIS

The issues in multilevel path analysis will be illustrated with a data set from a study by Schijf and Dronkers (1991). They analyzed data from 1377 pupils in 58 schools, and from this we have the following pupil-level variables: father's occupational status *focc*, father's education *feduc*, mother's education *meduc*, pupil sex *sex*, the result of the

GALO school achievement test *galo*, and the teacher's advice about secondary education *advice*. At the school level we have one global variable: the school's denomination *denom*. Denomination is coded 1 = Protestant, 2 = Nondenominational, 3 = Catholic (categories based on optimal scaling). The research question is whether the school's denomination affects the GALO score and the teacher's advice, after the other variables have been accounted for.[1]

We can use a sequence of multilevel regression models to answer this question. The advantage of a path model is that we can specify one model that describes all hypothesized relations between independent, intervening, and dependent variables. However, we have multilevel data, with one variable at the school level, so we must use a multilevel model to analyze these data.

Figure 15.1 shows part of the school data. The GALO data illustrate several problems that occur in practice. First, the school variable *denom* does not vary within schools. This means that it has an intraclass correlation of 1.00, and must be included only in the between model. In this specific example, there is another problematic



| | school | sex | galo | advice | feduc | meduc | focc | denom |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 2 | 78 | 1 | 1 | 1 | 2 | 2 |
| 2 | 1 | 1 | 104 | 4 | 4 | 3 | 4 | 2 |
| 3 | 1 | 2 | 93 | 2 | 1 | 1 | 2 | 2 |
| 4 | 1 | 1 | 114 | 4 | 2 | 1 | 5 | 2 |
| 5 | 1 | 1 | 95 | 2 | 2 | 2 | 5 | 2 |
| 6 | 1 | 1 | 98 | 2 | 1 | 1 | 2 | 2 |
| 7 | 1 | 1 | 114 | 4 | 1 | 3 | 1 | 2 |
| 8 | 1 | 1 | 79 | 2 | 1 | 1 | 999 | 2 |
| 9 | 1 | 2 | 84 | 2 | 1 | 1 | 2 | 2 |
| 10 | 1 | 2 | 101 | 2 | 1 | 1 | 2 | 2 |
| 11 | 1 | 2 | 99 | 2 | 3 | 4 | 2 | 2 |
| 12 | 1 | 2 | 102 | 2 | 5 | 5 | 4 | 2 |
| 13 | 1 | 2 | 86 | 2 | 1 | 1 | 4 | 2 |
| 14 | 2 | 1 | 110 | 2 | 4 | 3 | 2 | 3 |
| 15 | 2 | 2 | 111 | 4 | 5 | 3 | 6 | 3 |
| 16 | 2 | 1 | 100 | 2 | 4 | 5 | 3 | 3 |
| 17 | 2 | 1 | 79 | 0 | 4 | 2 | 2 | 3 |
| 18 | 2 | 2 | 111 | 4 | 2 | 1 | 3 | 3 |

*Figure 15.1* Part of data file for school example.

[1] The data were collected in 1971. The same data were analyzed in Hox (2002) using MUML. The current model was suggested by Gustafsson and Stahl (1999).

variable, *sex*, which is pupil sex. This variable turns out to have an intraclass correlation of .005, which is very small, and obviously not significant. This means that there is almost no variation between schools in the gender composition; all schools have about the same proportion of girls and boys. As a result of this empirical finding, the variable *sex* can only be used at the pupil level, and must be omitted from the school level. The other variables have intraclass correlations that range from .14 for advice to .29 for father occupation. The analyses in this chapter have been carried out using Mplus, which has the option of declaring variables as existing only at the within or the between level. Since pupil sex turned out to have no significant covariances with any of the other variables in the analysis, it is completely removed from the analyses.

An additional problem is the occurrence of missing data, coded in the data file as 999. Schijf and Dronkers (1991) analyzed only the complete data ($N = 1377$) using a series of multilevel regression models, and Hox (2002) also analyzes only the complete data using MUML path analysis. Listwise deletion of incomplete cases assumes that the incomplete data are missing completely at random (MCAR). In the GALO data, the largest proportion of incomplete data is in the variable father occupation, which is missing in almost 8% of the cases. Missing father occupation may be the result of father being unemployed, which is likely to be correlated with education, and is also likely to affect the advice. In fact, missingness on the *focc* variable is related to both *feduc* and *advice*, with a missing code for father occupation related to a lower level of father education and a lower value for advice. Evidently, an analysis method that assumes missing completely at random is likely to result in biased estimates. Full maximum likelihood analysis of incomplete data assumes missing at random (MAR), a much weaker assumption. (The distinction between MCAR and MAR is discussed in more detail in Chapter 5 in the context of dropout in longitudinal data.) Thus, all analyses are carried out using maximum likelihood on all cases, using robust estimation (MLR). In this analysis 1559 pupils are included, who are in 58 schools.

### 15.1.1 Preliminaries: Separate analysis for each level

The pooled within-groups covariance matrix $\mathbf{S}_{PW}$ is in this case based on 1559 − 58 = 1501 cases. Since $\mathbf{S}_{PW}$ is an unbiased estimate of $\mathbf{\Sigma}_W$, we can use it for a preliminary analysis at the pupil level. The problematic variable *sex* is removed completely from the model, since it has no significant relations with any of the other variables. As there are incomplete data, the sample size is undefined. As an approximation, we specify a sample size based on the complete data: $N = 1319$ (1377 − 58). Since robust estimation requires raw data, the preliminary analyses on $\mathbf{S}_{PW}$ use normal ML estimation.

Figure 15.2 depicts the pupil-level model, which contains one latent variable, *ses*, measured by the observed variables *focc*, *feduc*, and *meduc*.

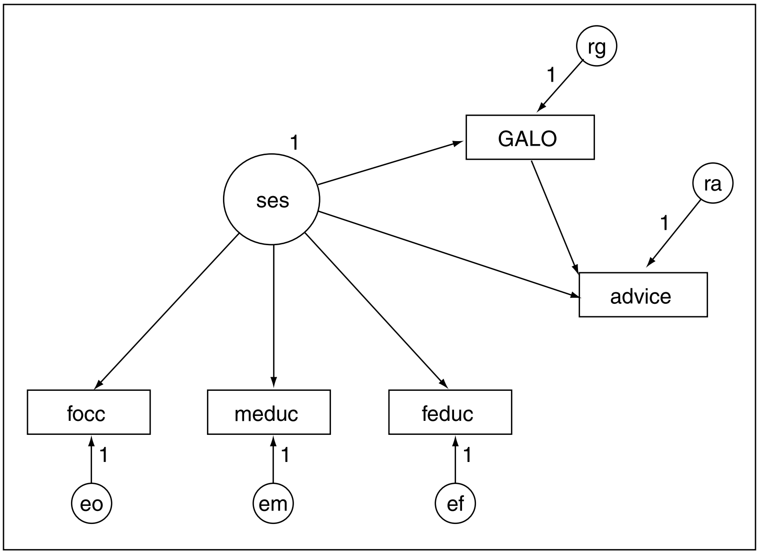


*Figure 15.2* Initial pupil-level path diagram.

The analysis of the pupil-level model on $\mathbf{S}_{PW}$ only gives a chi-square of 15.3, with *df* = 4 and $p < .01$. The goodness-of-fit indices are good: CFI = 1.00, TLI = .99, RMSEA = .05. Modification indices suggest adding a covariance between the residuals of father occupation and father education. If this residual covariance is added, we obtain a chi-square of 3.5, with *df* = 3 and $p = .32$. The goodness-of-fit indices are excellent: CFI = 1.00, TLI = 1.00, RMSEA = .01. This model is accepted.

The next step is specifying a school-level model. When maximum likelihood estimation is used, we can estimate a saturated model for both the within and the between level, which will provide the maximum likelihood estimate for $\mathbf{\Sigma}_B$.[2] We start the analysis of the estimated between-groups matrix $\hat{\mathbf{\Sigma}}_B$ by specifying the initial pupil-level model as depicted in Figure 15.2, setting the sample size at 58. The school-level variable denomination is used as a predictor variable for both GALO and advice. This model is rejected: chi-square is 63.3, with *df* = 6 and $p < .01$. The goodness-of-fit indices also indicate bad fit: CFI = .92, TLI = .79, RMSEA = .41. In addition, the estimate of the residual variance of father education is negative, and the effect of denomination on advice is not significant ($p = .64$). Further pruning of the model leads to the model depicted in Figure 15.3. This model still does not fit well: chi-square

[2] With Mplus, estimates for $\mathbf{\Sigma}_W$ and $\mathbf{\Sigma}_B$ are produced as part of the sample statistics output.

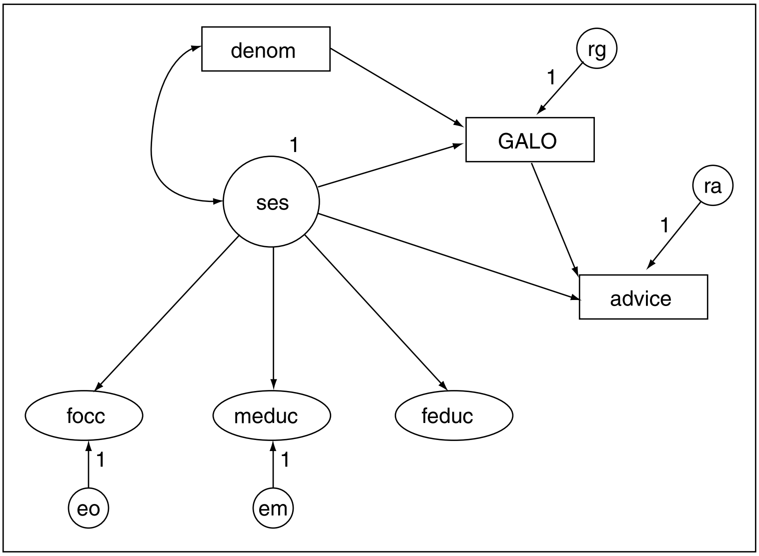


*Figure 15.3* Final school-level path diagram, *ses* as latent variable.

is 65.6, with *df* = 8 and *p* < .01. The goodness-of-fit indices also indicate bad fit: CFI = .92, TLI = .84, RMSEA = .35. There are no large modification indices, so there are no obvious ways to improve the school-level model.

Observation of the school-level correlation matrix shows that at the school level father education and mother education have a high correlation, while the correlations with father occupation are much lower. Also, the covariances of father occupation and father education with other variables appear different. This indicates that assuming a latent variable *ses* at the school level may be wrong. An entirely different way to model the effect of these variables on GALO and advice is to use a model with only observed variables. The initial path diagram for such a model is shown in Figure 15.4.

The model depicted in Figure 15.4 is a saturated model, and therefore cannot fail to fit. It turns out that the effects of denomination and father occupation on advice are not significant, so the final model becomes the path diagram in Figure 15.5.

This final school-level model fits well: chi-square is 2.1, with *df* = 2 and *p* = .34. The goodness-of-fit indices also indicate good fit: CFI/TLI = 1.00, RMSEA = .04. The SES model in Figure 15.4 and the regression model in Figure 15.6 are not nested, but they can be compared using the information criteria AIC and BIC. For the SES model AIC = 148.03 and BIC = 172.76, and for the regression model AIC = 78.56 and BIC = 97.11. Both the AIC and the BIC indicate a preference for the regression model.

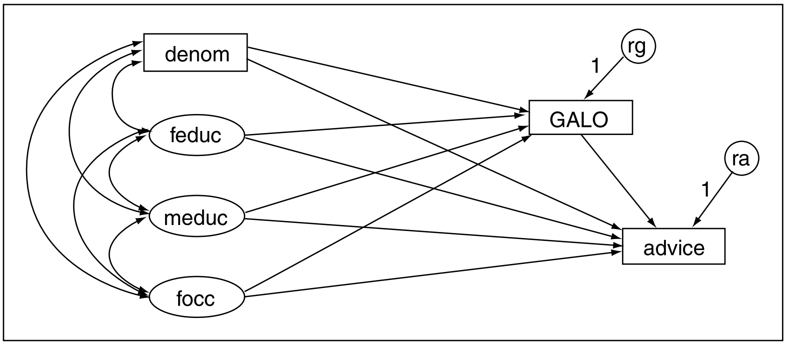


*Figure 15.4* Initial school-level path diagram, regression model.

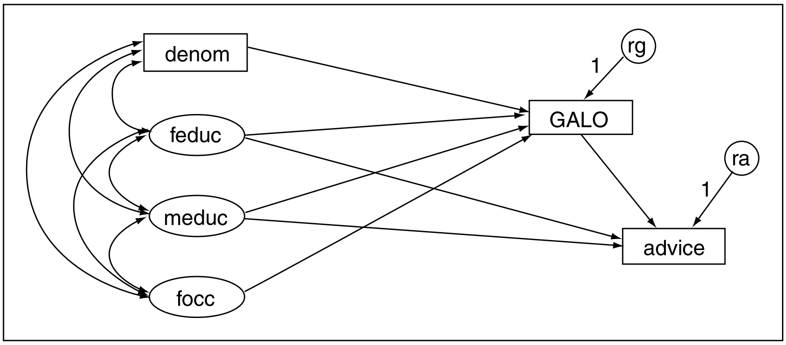


*Figure 15.5* Final school-level path diagram, regression model.

### 15.1.2 Putting it together: Two-level analysis

The preliminary analyses give some indication of what to expect when the models are combined into a two-level model, with simultaneous estimation at both levels. There will be differences. First, because the two-level analysis is simultaneous, misspecifications at one level will also affect the other level. This has a positive side as well. The estimated between-groups covariance matrix used earlier is estimated using a saturated model for the within-groups part. If we have a well-fitting and parsimonious within model, the between model is more stable. Second, the preliminary analyses are based on maximum likelihood estimates of the corresponding population covariance matrices, but in the presence of incomplete data the stated sample size is only an approximation.

When the individual-level model is combined with a school-level model in a simultaneous analysis, it turns out that the model with latent variable *ses* at the school level fits better than the model with regressions on observed variables only. Table 15.1 presents the chi-squares and fit indices for several different school-level models, all fitted in combination with the within-schools model established earlier, and the fit indices CFI, TLI, and RMSEA calculated by hand specifically for the school level.

*Table 15.1* Fit indices for several school-level models

| Model | $\chi^2$ | *df* | *p* | CFI | TLI | RMSEA | AIC* | BIC* |
|---|---|---|---|---|---|---|---|---|
| Indep. | 577.7 | 18 | .00 | – | – | – | 5560 | 5684 |
| SES | 11.0 | 11 | .45 | 1.00 | 1.00 | .00 | 5124 | 5279 |
| Regress. | 22.9 | 8 | .00 | .97 | .94 | .03 | 5140 | 5311 |

* AIC and BIC are 25,000 for legibility.

Thus, the final two-level model is for the pupil level as depicted in Figure 15.6, and the school-level model is as depicted earlier in Figure 15.4. The combined model fits well: chi-square = 14.9, *df* = 11, *p* = .45, CFI/TLI = 1.00, RMSEA = .02.

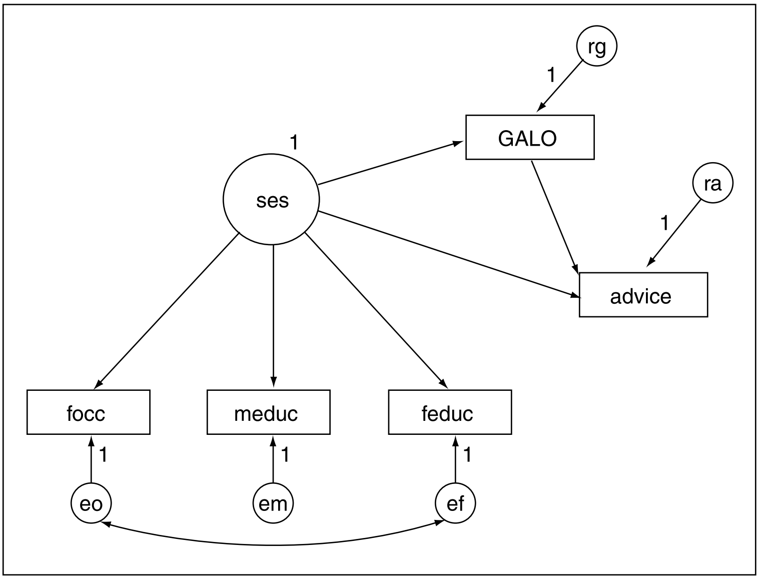


*Figure 15.6* Final pupil-level model.

Since we have a latent variable *ses* at both the individual and the school level, the question is relevant if we can constrain the loadings of father occupation, mother education, and father education to be identical across the two levels. A model with equality constraints for these loadings across the two levels fits quite well: chi-square = 20.6, *df* = 14, *p* = .11, CFI/TLI = 1.00, RMSEA = .02. Since the latent variable *ses* is now defined identically at both levels, we can formally compare its variance at the within and the between level. The variance of SES is fixed at 1.00 at the individual level, and freely estimated as 0.62 at the school level. The intraclass correlation for the latent variable SES is .38, which is considerably higher than the ICC for its constituent observed variables father occupation, mother education, and father education. This is typical, since measurement error in the observed variables tends to end up at the lowest level (Muthén, 1991b).

Table 15.2 presents the path coefficients and corresponding standard errors. There are strong school-level effects on the GALO score and on the advice. The school-level variable *denomination* turns out to have an effect only on the school-level GALO test score.

*Table 15.2* Path coefficients for final GALO model

| Effect on | GALO – pupil | Advice – pupil | GALO – school | Advice – school |
|---|---|---|---|---|
| Effect from: | | | | |
| SES | 0.43 (.04) | 0.12 (.02) | 0.52 (.09) | 0.31 (.06) |
| GALO | – | 0.85 (.03) | – | 0.54 (.11) |
| Denom. | – | – | 0.24 (.08) | Not significant |

In the multilevel regression analyses presented by Schijf and Dronkers (1991), denomination has a significant effect on both the teachers' advice and on the GALO test score. The path model presented here shows that the main influence is through the GALO test score; the different types of advice given by teachers in schools of different denominations are apparently the result of differences in GALO test scores between such schools. This is precisely the kind of result that a sequence of separate regression analyses cannot show.

Figure 15.6 also shows that SES has a school-level effect on the variables GALO and advice. The interpretation is *not* that some schools simply happen to have more high SES pupils and therefore perform better; sampling differences between schools in SES composition are accounted for in the pupil model that is also fitted for the school-level covariances. Instead, the substantive interpretation of the school-level results must be in terms of some kind of contextual or systematic selection effect. It appears that the concentration of high or low SES pupils has its own effect on the school career variables. It is interesting to note that, at the school level, the effect of the school average on the GALO test on the average advice is negative. This can be interpreted as a typical context effect, in which the GALO score is apparently interpreted differently by teachers if the overall score of a school on the test is high.

## 15.2 STATISTICAL AND SOFTWARE ISSUES IN MULTILEVEL FACTOR AND PATH MODELS

It should be noted that multilevel factor and path models differ from multilevel regression models, because in general they do not have random regression slopes. The variation and covariation at the group level is intercept variation. There are also in general no cross-level and interaction effects. In multilevel factor models, the group-level variation can properly be interpreted as group-level variance of the group means of the latent factors. In path analysis, the interpretation of group-level path coefficients is in terms of contextual effects, which are added to the individual effects. Inclusion of cross-level and other interaction terms in structural equation models is possible, but they are difficult to specify and estimate in most current SEM software (see Jöreskog & Yang, 1996; Marcoulides & Schumacker, 2001; Schumacker & Marcoulides, 1998).

The two-step direct estimation approach does not require special software. The Muthén pseudobalanced approach can be applied in any SEM software that supports multiple group analysis. However, both Muthén's pseudobalanced approach and direct estimation of the within and between covariance matrices have strong limitations. Direct estimation of the covariance matrices, followed by standard SEM analyses, ignores the fact that these covariances are themselves estimates, and that with incomplete data it is misleading to assign a single sample size to each matrix. The pseudobalanced approach, although consistent, is not very accurate. As discussed in the chapter on multilevel CFA, both simulation studies (Hox and Maas, 2001; Hox et al., 2010) and analytical work (Yuan and Hayashi, 2005) show that the degree of inaccuracy depends on the degree of unbalance, and that larger sample sizes do not improve the accuracy with unbalanced data.

Raudenbush and Sampson (1999b) advocate a different method for analyzing multilevel models with latent variables, using standard multilevel regression software. They represent measurement error at a separate 'variables' level (Raudenbush, Rowan, & Kang, 1991), a method described in the multivariate multilevel measurement models section in Chapter 10 of this book. The random regression coefficients at the second level are interpreted as latent variables or factors, indicated by the variables to which they are linked by sets of dummy variables. Using the means and covariances at the higher levels, path coefficients can be estimated with the corresponding standard errors. This approach, incorporated in the software HLM, can be used to estimate both factor and path models. The major advantages of their approach are that it can include random regression coefficients, and it works fine with incomplete data, which are difficult to handle in the pseudobalanced approach. The major disadvantage is the simplicity of the measurement model. The model requires that all the factor loadings are known; typically they are all set equal to one. This is a strong assumption, which implies that all observed variables that load on the same factor are measured on the same scale and have the same error variance. There are also no fit indices, so information on how well the factor or path model fits is not readily available.

A full maximum likelihood solution to the problem of multilevel factor and path analysis requires maximization of a complicated likelihood function. Bentler and Liang (2001) describe a method for maximizing the multilevel likelihood for both confirmatory factor and path models, which is implemented in the software EQS. LISREL (8.5 and later; du Toit & du Toit, 2001) includes a full maximum likelihood estimation procedure for multilevel confirmatory factor and path models, including an option to analyze incomplete data. The LISREL 8.5 user's manual (du Toit & du Toit, 2001) cautions that this procedure still has some problems; it frequently encounters convergence problems, and needs good starting values. Mplus (Muthén & Muthén, 1998–2007) includes several options for multilevel factor and path analysis. In addition to the pseudobalanced MUML approach, Mplus offers weighted least squares and full maximum likelihood estimation. Mplus also allows fitting multilevel structural equation models with random coefficients, regressions among latent variables varying at two different levels, and mixtures of continuous and ordered or dichotomous data. Rabe-Hesketh and Skrondal (2008) describe a very general software called GLLAMM (for Generalized Linear Latent And Mixed Models) that runs in the statistical package STATA. The program and manual are available free (Rabe-Hesketh, Skrondal, & Pickles, 2004), but the commercial package STATA is needed to run it. Like Mplus, GLLAMM can fit very general multilevel structural equation models, with no formal limit to the number of levels.

Many more complex types of multilevel path models with latent variables, for instance including random slopes and cross-level effects, can be estimated using Bayesian methods. Bayesian models for continuous data are described by Goldstein and Browne (2001) and Jedidi and Ansari (2001), and models for binary data by Ansari and Jedidi (2000). Many multilevel structural equation models can be estimated using Bayesian methods and the software BUGS (Spiegelhalter, 1994), but this requires an intimate knowledge of both structural equation modeling and Bayesian estimation methods. The software REALCOM (multilevel modeling for realistically complex data; Goldstein, Steele, Rasbash, & Charlton, 2007) also includes Bayesian methods for complex multilevel modeling.

It is clear that there have been many developments in multilevel factor and path analysis, but most require sophisticated and detailed statistical knowledge. I consider them beyond the scope of this book, especially since it will take time before these methods find their way in generally available and user-friendly software packages. Given its general availability in modern SEM software, maximum likelihood estimation is the preferred method. For non-normal data, weighted least squares is attractive because it is computationally faster, but this is available only in Mplus.

The analysis issues in multilevel path models, whether analyzed using Muthén's pseudobalanced approach, using directly estimated within- and between-groups covariance matrices, or full maximum likelihood estimation, are comparable to the issues in multilevel factor analysis. Thus, the recommendations given in Chapter 14 about inspecting goodness-of-fit indices separately for the distinct levels that exist in the data, and about the separate standardization, also apply to multilevel path models.

All approaches to multilevel factor and path analysis model only one single within-groups covariance matrix. In doing so, they implicitly assume that the within-groups covariances are homogeneous, that is, that all groups have the same within-groups covariance matrix. This is not necessarily the case. The effect of violating this assumption is currently unknown. Simulation studies on the assumption of homogeneous covariance matrices in MANOVA show that MANOVA is robust against moderate differences in the covariances, provided the group sizes are not too different (Stevens, 2009). Strongly different group sizes pose a problem in MANOVA. When larger variability exists in the smaller group sizes, the between-group variation is overestimated; when larger variability exists in the larger group sizes, the between-group variation is underestimated.

If we assume that the covariance matrices differ for different groups, one possible solution is to divide the groups into two or more separate sub-sets, with each sub-set having its own within-groups model. For instance, we may assume that within-groups covariances differ for male and female respondents. Or, in the situation where we have larger variances in small groups and vice versa, we may divide the data into a set of small and a set of large groups. Then we model a different within-groups model for each set of groups, and a common between-groups model. Mplus and GLLAMM allow latent groups, which is a different way to allow several different covariance matrices at the individual or the group level.

## APPENDIX

The global variable *denomination* does not exist at the individual level. If we disaggregate denomination to the individual level, we will find that this disaggregated variable is constant within each school, and that the variance and all the covariances with the other individual deviation scores are zero. In software that requires the same number of variables in both groups, this problem is solved by viewing the school-level variable denomination as a variable that is systematically missing in the pupil-level data. The trick is that the variable denomination is included in the (school-level) between-schools covariance matrix in the usual way. In the (pupil-level) within-schools covariance matrix, we include denomination as a variable with a variance equal to one and all covariances with other variables equal to zero. In the within-school models, there are no paths pointing to or from this observed variable. Subsequently, we constrain the residual error variance to 1.00 for this variable. Since this produces a perfect fit for this 'ghost' variable, inclusion of this variable has no influence on the within-school estimates or the overall chi-square. There is only one problem; a program like LISREL assumes that this variable represents a real observed variable, and will include it when it enumerates the degrees of freedom for the within-schools model. As a result, the *df* and therefore the *p*-values (and most fit indices) in the output are incorrect, and must be

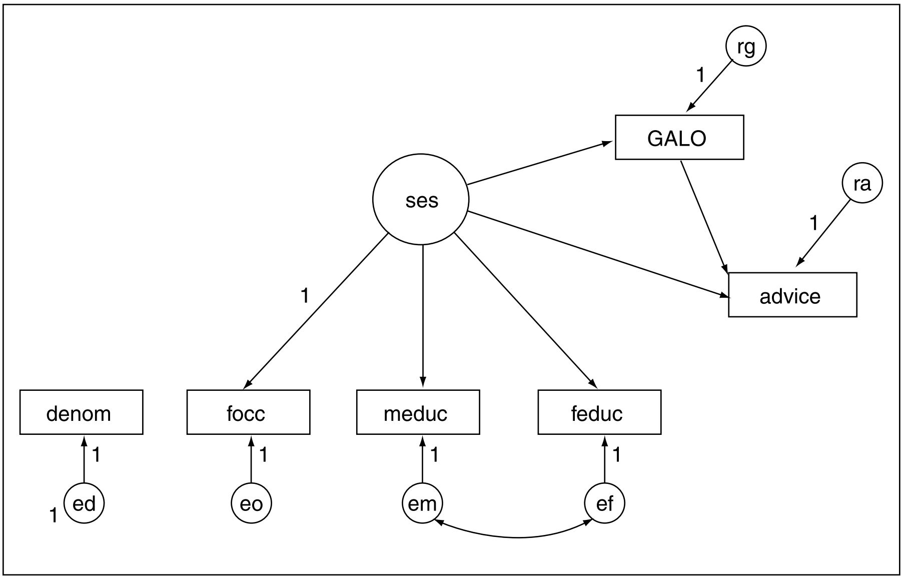


*Figure 15A.1* Including a school-level variable at the within level.

recalculated by hand (see Jöreskog & Sörbom, 1989). The *df* value is corrected by subtracting the number of zero covariances from the *df* calculated by the program. Again, some software can handle models with different numbers of observed variables in the various groups, which makes this kind of modeling much simpler.

Figure 15A.1 shows how a school-level variable like *denom* is handled in a program like LISREL using MUML. It is included in the pooled within-groups covariance matrix as a virtual variable with a variance of one, and the covariances with the other variables all set to zero. In the pupil-level model, which is the model for the pooled within-groups covariance matrix, the observed variable *denom* is modeled as independent from all other variables in the model, with an error term *ed* that has a variance fixed at one. This of course exactly models the variance of one and the covariances of zero that were added to the pooled within-groups covariance matrix. The problem is that a program like LISREL counts an extra variance plus five covariances in the input matrix, which means that it counts an extra six degrees of freedom. Since these nonexistent variance and covariances values are necessarily estimated perfectly, the fit of the model appears spuriously good. For this reason, the degrees of freedom *df* must be corrected by subtracting 6, and the *p*-value and all goodness-of-fit indices that depend on the degrees of freedom must also be corrected.

# 16

# Latent Curve Models

An interesting structural equation model for fixed occasion panel data is the latent curve model (LCM). This model has been applied mainly to developmental or growth data, hence the usual name 'latent growth model'. In the latent curve model, the time or measurement occasion variable is defined in the measurement model of the latent factors. For instance, in a linear growth model, consecutive measurements are modeled by a latent variable for the intercept of the growth curve, and a second latent variable for the slope of the curve.

Figure 16.1 shows the path diagram of a simple latent curve model for panel data with five occasions, and one time-independent explanatory variable $Z$. In Figure 16.1, $Y_0$, $Y_1$, $Y_2$, $Y_3$, and $Y_4$ are the observations of the response variable at the five consecutive time points. In the latent curve model, the expected score at time point zero is modeled by a latent *intercept* factor. The intercept is constant over time, which is

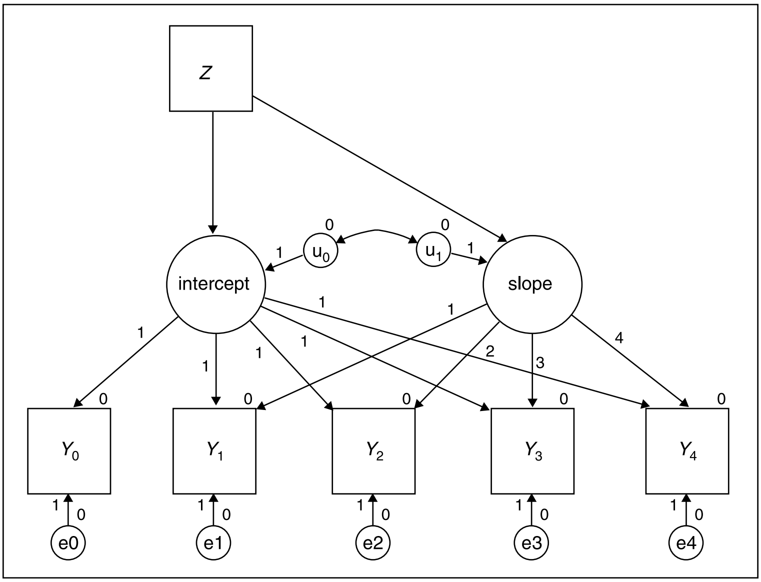


*Figure 16.1* Latent curve model for five occasions.

modeled by constraining the loadings of all time points on the intercept factor to be equal to one. The latent slope factor is the slope of a linear curve, modeled by constraining the loadings of the five time points on this factor to be equal to 0, 1, 2, 3, and 4 respectively. Following the usual custom in the graphical model presentation in SEM, the one path that is constrained to zero is not drawn. Obviously, a quadratic trend would be specified by a third latent variable, with successive loadings constrained to be equal to 0, 1, 4, 9, and 16. What is not immediately obvious from the path diagram in Figure 16.1 is that the latent curve model *must* include the intercepts of the observed variables and the means of the factors in the model. As a result, the regression equations that predict the observed variables from the latent factors, depicted by the single-headed arrows towards the observed variables in Figure 16.1, also contain terms for the intercept.

In the latent curve model, the intercepts of the response variable at the five time points are all constrained to zero, and as a result, the mean of the intercept factor is an estimate of the common intercept. In Figure 16.1, which uses the AMOS representation (Arbuckle, 2007), this is visible in the zeros placed close to the latent and observed variables; these indicate means and intercepts that are constrained to zero.

The successive loadings for the slope factor define the slope as the linear trend over time (the first path from the slope factor to variable $Y_0$, which is equal to zero, is omitted from the diagram). The mean of the slope factor is an estimate of the common slope (see Duncan, Duncan, & Strycker, 2006; Meredith & Tisak, 1990; Muthén, 1991a; Willett & Sayer, 1994). Individual deviations from the common intercept are modeled by the variance of the intercept factor, and individual deviations in the slope of the curve are modeled by the variance of the slope factor. Both the intercept and the slope factor can be modeled by a path model including explanatory variables, in our example the one explanatory variable $Z$.

The latent curve model is a random coefficient model for change over time, completely equivalent to the multilevel regression model for longitudinal data that is described in Chapter 5. To clarify the relationship between the two models, we write the equations for both specifications. In the multilevel linear growth model, the model described by Figure 16.1 can be expressed as a multilevel regression model, with at the lowest level, the occasion level:

$$Y_{ti} = \pi_{0i} + \pi_{1i}T_{ti} + e_{ti}, \tag{16.1}$$

where $T_{ti}$ is an indicator for the occasions, which is set to 0, 1, 2, 3, 4 to indicate the five occasions, with subscript $t$ indicating occasions, and subscript $i$ the individuals. . At the second level, the individual level, we have:

$$\pi_{0i} = \beta_{00} + \beta_{01}Z_i + u_{0i} \tag{16.2}$$

$$\pi_{1i} = \beta_{10} + \beta_{11}Z_i + u_{1i} \tag{16.3}$$

By substitution, we get the single equation model:

$$Y_{ti} = \beta_{00} + \beta_{10}T_{ti} + \beta_{01}Z_i + \beta_{11}Z_iT_{ti} + u_{1i}T_{ti} + u_{0i} + e_{ti} \tag{16.4}$$

In a typical SEM notation, we can express the path model in Figure 16.1 as:

$$Y_{ti} = \lambda_{0t}\ intercept_i + \lambda_{1t}\ slope_i + e_{ti} \tag{16.5}$$

where $\lambda_{0t}$ are the factor loadings for the intercept factor, and $\lambda_{1t}$ are the factor loadings for the slope factor.

Note the similarity between the equations 16.5 and 16.1. In both cases, we model an outcome variable that varies across times *t* and individuals *i*. In equation 16.1, we have the intercept term $\pi_{0i}$, which varies across individuals. In equation 16.5, we have a latent intercept factor, which varies across individuals, and is multiplied by the factor loadings $\lambda_{0t}$ to predict the $Y_{tj}$. Since the factor loadings $\lambda_{0t}$ are all constrained to be equal to one, they can be left out of equation 16.5, and we see that the intercept factor in equation 16.5 is indeed equivalent to the regression coefficient $\pi_{0i}$ in equation 16.1. Next, in equation 16.1, we have the slope term $\pi_{1i}$, which varies across individuals, and is multiplied by the 0, . . . , 4 values for the occasion indicator $T_{ti}$. In equation 16.5, we have a latent slope factor, which varies across individuals, and gets multiplied by the factor loadings $\lambda_{0t}$ to predict the $Y_{tj}$. Since the factor loadings $\lambda_{1t}$ are set to 0, . . . , 4, we see that the slope factor in equation 16.5 is indeed equivalent to the regression coefficient $\pi_{1i}$ in equation 16.1. Therefore, the fixed factor loadings for the slope factor play the role of the time variable $T_{ti}$ in the multilevel regression model, and the slope factor plays the role of the slope coefficient $\pi_{1i}$ in the multilevel regression model. The random regression coefficients in the multilevel regression model are equivalent to the latent variables in the latent curve model.

In a manner completely analogous to the second-level equations 16.2 and 16.3 in the multilevel regression model, we can predict the intercept and the slope factor using the time-independent variable *Z*. For these equations, using for consistency the same symbols for the regression coefficients, we have:

$$intercept_i = \beta_{00} + \beta_{01}Z_i + u_{0i} \tag{16.6}$$
$$slope_j = \beta_{10} + \beta_{11}Z_i + u_{1i} \tag{16.7}$$

which leads to a combined equation:

$$Y_{ti} = \beta_{00} + \beta_{10}\ \lambda_{1t} + \beta_{01}Z_i + \beta_{11}Z_i\ \lambda_{1t} + u_{1i}\ \lambda_{1t} + u_{0i} + e_{ti} \tag{16.8}$$

Keeping in mind that the factor loadings 0, . . . , 4 in $\lambda_{1t}$ play the role of the occasion indicator variable in $T_t$, we see that the multilevel regression model and the latent curve model are indeed equivalent. The only difference so far is that multilevel regression analysis generally assumes one common variance for the lowest-level errors $e_{ti}$, while structural equation analysis typically estimates different residual error variances for all observed variables. However, this is easily solved by imposing a constraint on the latent curve model that the variances for $e_0, \ldots, e_4$ are all equal. If we impose this constraint, we have indeed the same model. Similarly, we can replace the single lowest-level error term in a multilevel regression by five error terms connected with five dummy variables, one for each time point (see for details the multivariate multilevel model presented in Chapter 10). Full maximum likelihood estimation, using either approach, should give essentially the same results.

## 16.1 EXAMPLE OF LATENT CURVE MODELING

The longitudinal GPA data from Chapter 5 are used again, with a standard latent curve model as in Figure 16.1 applied to the data. The example data are a longitudinal data set, with the data coming from 200 college students. The students' grade point average (GPA) was recorded for six successive semesters. At the same time, it was recorded whether the student held a job in that semester, and for how many hours. This is recorded in a variable job. In this example, we also use the student variables high school GPA and sex (0 = male, 1 = female).

In a statistical package such as SPSS or SAS, such data are typically stored with the students defining the cases, and the repeated measurements as a series of variables, such as GPA1, GPA2, . . . , GPA6, and JOB1, JOB2, . . . , JOB6. As explained in Chapter 5, most multilevel regression software requires a different data structure. However, latent curve analysis views the successive time points as multivariate outcome variables, and thus we can use the data file as it is. We start with a model that includes only the linear trend over time. Figure 16.2 shows the path diagram for this model.

The model in Figure 16.2 is equivalent to a multilevel regression model with a linear predictor coded 0, . . . , 5 for the successive occasions, and a random intercept and slope at the student level. To make the models completely equivalent, the error variances of the residual errors $e_1, \ldots, e_6$ for the successive occasions are all constrained to be equal to $e$. In the graphical model in Figure 16.2 this is symbolized by the letter $e$ next to each residual error variable. The means of the intercept and slope factor are freely estimated; all other means and intercepts in the model are constrained to zero, which is symbolized by placing a zero next to the constrained variable. The mean of the intercept is freely estimated as 2.60, and the mean of the slope is estimated as

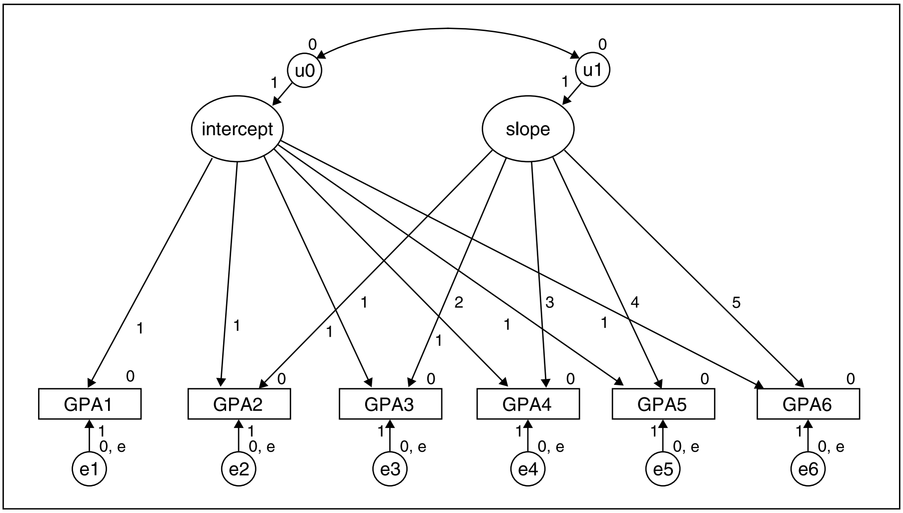


*Figure 16.2* Path diagram for linear model GPA example data.

0.11. This is identical to the estimates of the intercept and slope in the (fixed effects) multilevel regression model in Table 5.3 in Chapter 5.

For simplicity, we omit the time-varying *job* variable for the moment, and start with specifying a latent curve model using only the six *GPA* scores, and the time-independent (student level) variables *high school GPA* and *student sex*. The path diagram, which includes the unstandardized parameter estimates obtained by standard SEM estimation, is shown in Figure 16.3.

In the path diagram we see that in this model, which includes an interaction between the slope of the linear development over time and the student's sex, the average slope over time is 0.06. The slope variance in the figure is given to two decimal places as 0.00; in the text output it is given as 0.004, with standard error .001. This is identical to the estimates in the similar multilevel regression model presented in Table 5.4 in Chapter 5.

The SEM analysis of the latent curve model gives us some information that is not available in the equivalent multilevel analyses. The fit indices produced by the SEM software tell us that the models depicted in Figures 16.2 and 16.3 do not describe the data well. The model in Figure 16.2 has a chi-square of 190.8 ($df = 21$, $p < .001$) and an RMSEA fit index of .20, and the model in Figure 16.3 has a chi-square of 195.3 ($df = 30$, $p < .001$) and an RMSEA fit index of .17. The SEM analysis also provides us with diagnostic information of the locus of the fit problem. The program output contains so-called *modification indices* that point out constraints that significantly

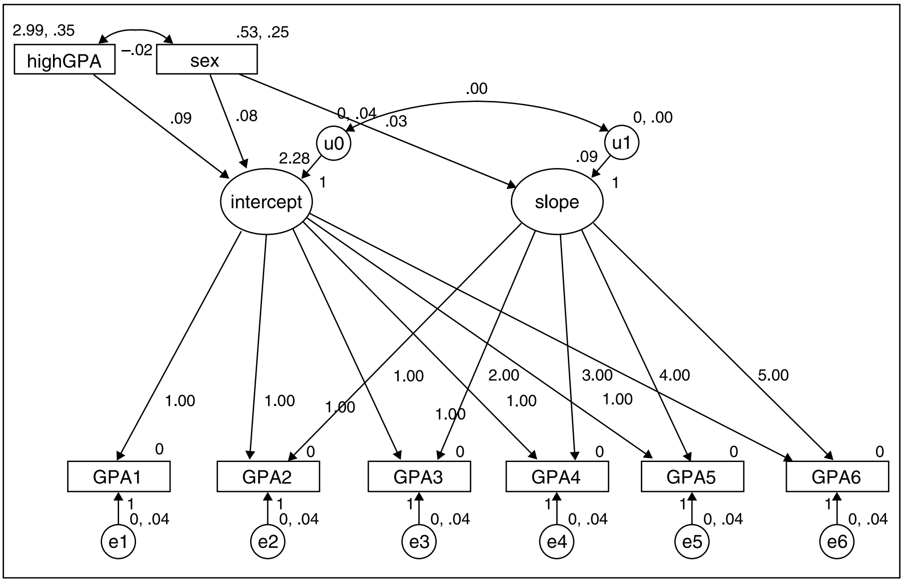


*Figure 16.3* Path diagram and parameter estimates for linear curve model with two predictors.

decrease the fit of the model. All large modification indices indicate that the constraint of equal error variances for the residual errors $e_1, \ldots, e_6$ does not fit well, and that the implicit constraint of no correlations between the residual errors $e_1, \ldots, e_6$ does not fit well either. Presumably, the multilevel regression models presented in Chapter 5 also have these problems. Since in Chapter 5 we did not carry out a residuals analysis or some other procedure to check for model misspecifications, we do not have any information about model fit. In SEM, we do have such information; it is automatically provided by all current SEM software. If we remove the equality constraint on the residual errors, the model fit becomes much better, as indicated by a chi-square of 47.8 ($df = 25$, $p = .01$) and an RMSEA fit index of .07. Allowing correlated errors between the two first measurement occasions improves the fit to a chi-square of 42.7 ($df = 24$, $p = .01$) and an RMSEA of .06. Since the other estimates do not change much because of these modifications, the last model is accepted.

To bring the time varying variable *job status* into the model, we have several choices. Equivalent to the multilevel regression models for these data, which are treated in Chapter 5, we can add the variables $job_1, \ldots, job_6$ as explanatory variables to the model. These predict the outcomes $GPA_1, \ldots, GPA_6$, and since the multilevel regression model estimates only one single regression for the effect of *job status* on *GPA*, we must add equality constraints for these regression coefficients if we want to estimate exactly the same model.

The path diagram for this model is given in Figure 16.4. Note the zeros that indicate means and intercepts constrained to zero. The variances of these variables are not constrained, which is visible in the diagram because there are no constraints visible next to the zeros. . The common regression coefficient for *job status* on the *GPA* is estimated as –0.12 (s.e. .01), which is close to the multilevel regression estimates in Table 5.4. However, the model including all the job status variables does not fit well, with a chi-square of 202.1 (*df* = 71, *p* < .001) and an RMSEA of .10. There are no large modification indices, which indicates that there is no single model modification that substantially improves the model. We probably need many small modifications to make the model fit better.

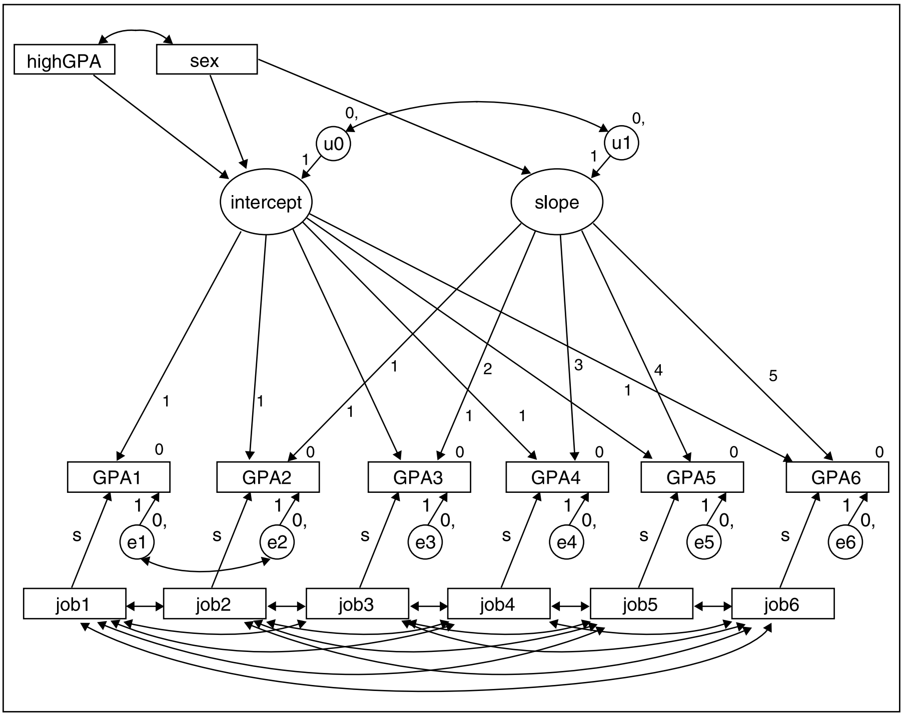


*Figure 16.4* Path diagram for GPA example, including effects for job status.

An advantage of latent curve analysis over multilevel regression analysis of repeated measures is that it can be used to analyze structures that are more complex.

For instance, we may attempt to model the changes in hours spent on a job using a second latent curve model. The path diagram for the latent curve model for *job status* at the six time points is presented in Figure 16.5.

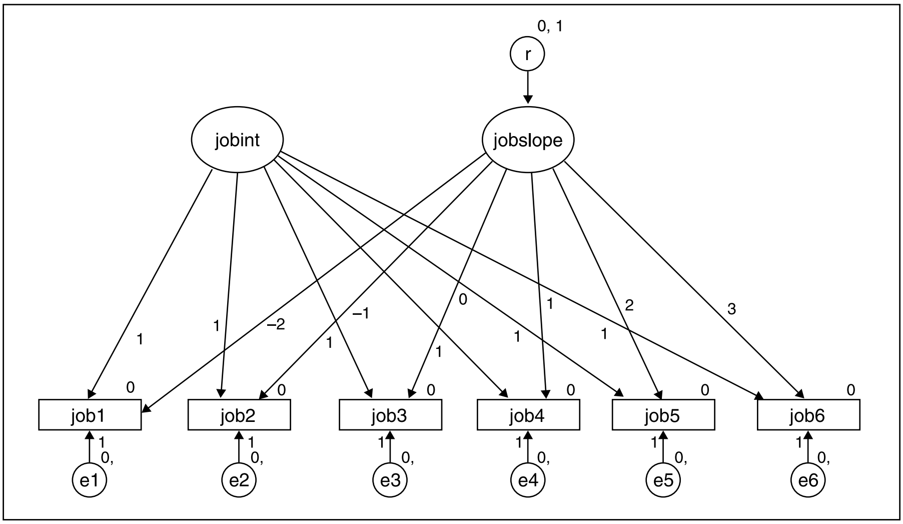


*Figure 16.5* Latent curve model for job status.

Figure 16.5 has some features that merit discussion. To ensure that the variance of the job slope factor is positive, the variance is modeled by an error term *r* with the variance set at 1, and the path to *jobslope* estimated. In this specific case, the more usual procedure of setting the path at 1 and estimating the error variances resulted in negative variance estimates. The model specification in Figure 16.6 leads to a latent curve model that fits quite well, with a chi-square of 17.8 ($df = 17$, $p = .40$) and an RMSEA of .02. All estimates in this model are acceptable. A powerful feature of structural equation modeling, compared to standard multilevel regression models, is that the models can be combined into one large model for change of both job status and GPA over time. Figure 16.6 shows an example.

The model depicted by the path diagram in Figure 16.6 has a moderate fit. The chi-square is 166.0 ($df = 85$, $p < .001$) and the RMSEA is .07. The AIC for the model in Figure 16.5, which is equivalent to a multilevel regression model, is 298.3. In comparison, the AIC for the model in Figure 16.7, which is *not* equivalent to a multilevel regression model, is 243.1. Although the complex latent curve model does not show an extremely good fit, it fits better than the related multilevel regression model.

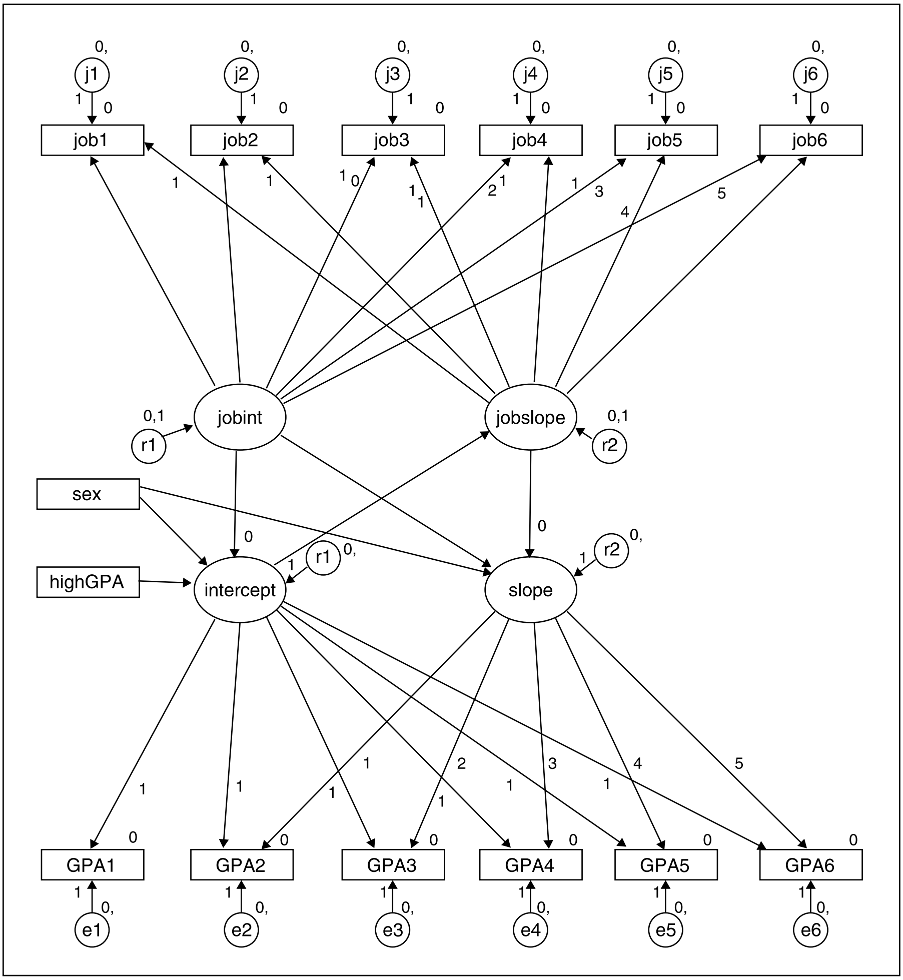


*Figure 16.6* Path diagram for change in job status and GPA over time.

Figure 16.6 also illustrates that for complicated models with constraints on intercepts and variances, a path diagram quickly becomes cluttered and difficult to read. At some point, presenting the model by describing a sequence of equations becomes simpler. Table 16.1 presents the estimates for the regression weights for the predictor variables *sex* and *high school GPA*, and the intercepts and slopes.

Figure 16.7 presents the same information, but now as standardized path coefficients with only the structural part of the path diagram shown.

Figure 16.7 shows results similar to the results obtained with the multilevel

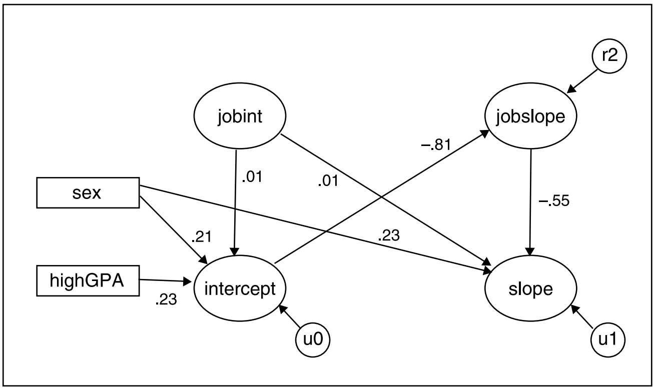


*Figure 16.7* Standardized path coefficients for structural model in Figure 16.6.

*Table 16.1* Path coefficients for structural model in Figure 16.7

| Predictor | Job slope (s.e.) | GPA interc. (s.e.) | GPA slope (s.e.) |
|---|---|---|---|
| Sex | | 0.07 (.03) | 0.02 (.01) |
| High sch. GPA | | 0.07 (.02) | |
| Job intercept | | 1.06 (.04) | 0.03 (.01) |
| Job slope | | | –0.46 (.11) |
| GPA intercept | –0.29 (.06) | | |

regression analyses in Chapter 5. Females have a higher GPA to begin with, and their GPA increases over the years at a faster rate than the male students do. The relations between the intercepts and slopes in Figure 16.7 show the mutual effects of changes over time in both job status and GPA. Initial job status has virtually no effect. Changes in job status, as reflected in the slope for job status, have a negative effect on the GPA. If the job status changes in the direction of spending more time on the job, the overall increase in GPA ends, and in fact can become negative. There is also an effect of initial GPA on job status: Students with a high initial GPA increase their job workload less than other students do.

## 16.2 A COMPARISON OF MULTILEVEL REGRESSION ANALYSIS AND LATENT CURVE MODELING

When equivalent multilevel regression analysis and latent curve modeling are applied to the same data set, the results are identical (Chou, Bentler, & Pentz, 2000). Plewis (2001) compares three different approaches to longitudinal data, including multilevel regression and latent curve models. Using empirical examples, he concludes that multilevel and latent curve models are very useful when testing interesting hypotheses about longitudinal data, for which they share many strengths and limitations.

A clear advantage of multilevel regression analysis is that adding more levels is straightforward. Modeling development over time of pupils nested within classes, nested in schools, is a simple procedure when multilevel regression is used, provided the analysis software can deal with more than three levels. When latent curve models and SEM software are used, adding a group level is possible (see Muthén, 1997). Adding a fourth level is virtually impossible. Multilevel regression also allows varying relationships at different levels, and modeling this variation by cross-level interactions with explanatory variables at the higher levels. In the SEM context, only the software Mplus (Muthén & Muthén, 1998–2007) and GLLAMM (Rabe-Hesketh et al., 2004) can deal with random slopes and cross-level interactions.

As remarked earlier in Chapter 5, multilevel regression copes automatically with missing data because of panel dropout. Since there is no requirement that each person has the same number of measurements, or even that the measures are taken on the same occasions, multilevel regression works very well on incomplete data. The latent curve model is a fixed occasions model. If different respondents are measured on different occasions, the latent curve model can deal with this only by specifying paths for all possible measurement occasions that occur in the data set, and regarding individuals observed on different measurement occasions as instances of incomplete data. Modern SEM software can estimate model parameters using maximum likelihood estimation on incomplete data (Arbuckle, 1996, 2007; du Toit & du Toit, 2001; Muthén & Muthén, 1998–2007), but when there are many and varying time points, the setup becomes complicated, and the estimation procedure may have convergence problems.

Latent curve models estimated with SEM software, on the other hand, have the advantage that it is straightforward to embed them in more complex path models. For instance, in latent growth methodology it is simple to specify a path model where the slope factor is itself a predictor of some outcome. This represents a hypothesis that the rate of growth is a predictor of some outcome variable. An example of such a path model was given in the previous section, where the rate of change in the latent slope for job status is a predictor for the rate of change indicated by the GPA slope factor. This model combines two latent curve models in one larger model, to investigate whether the rate of change in one process depends on the rate of change in a second process.

Although the program HLM allows some limited means of modeling latent variables and path models (Raudenbush et al., 2000), this kind of hypothesis is difficult to model in standard multilevel software. Hoeksma and Knol (2001) present an example of a model where the slope factor is a predictor of an outcome variable. Their discussion makes clear that these models can be specified in the multilevel regression framework, but also that it is difficult and leads to complicated software setups. Using the latent curve approach, this model is a straightforward extension of the basic latent curve model.

In the SEM latent growth curve approach, it is also simple to allow for different errors or correlated errors over time, which is possible in multilevel regression analysis, but more difficult to set up in the current software. A second interesting extension of the latent curve model is adding a measurement model for the variable that is measured over time or for the explanatory variables. To add a measurement model to the variable that is measured over time, it is indicated by a set of observed variables, and the variable that is modeled using the latent curve defined by the intercept and slope factors is then itself a latent variable. A final advantage of the latent curve model is that standard SEM software provides information on goodness of fit, and suggests model changes that improve the fit.

Finally, it should be noted that in latent curve models, the variances of the intercept and slope factors are important. The general usage in SEM is to test variances using the Wald test. As discussed in Chapter 3, using the Wald test for variances is not optimal, and the chi-square deviance difference test is much better. The same holds for latent curve models in SEM. As explained by Berkhof and Snijders (2001), the resulting *p*-value must be divided by two (see Chapter 3). Stoel et al. (2006) discuss this in the context of latent curve modeling using SEM.

# References

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, *22* (1), 47–76.

Afshartous, D. (1995). *Determination of sample size for multilevel model design*. Paper, AERA Conference, San Francisco, April 18–22, 1995.

Agresti, A. (1984). *Analysis of ordinal categorical data*. New York: Wiley.

Agresti, A., Booth, J. G., Hobart, J. P., & Caffo, B. (2000). Random effects modeling of categorical response data. *Sociological Methodology*, *30*, 27–80.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interaction*. Newbury Park, CA: Sage.

Aitkin, M., Anderson. D., Francis, B., & Hinde, J. (1989). *Statistical modelling in GLIM*. Oxford, UK: Clarendon Press.

Akaike, H. (1987). Factor analysis and the AIC. *Psychometrika*, *52*, 317–332.

Alba, R. D., & Logan, J. R. (1992). Analyzing locational attainments: Constructing individual-level regression models using aggregate data. *Sociological Methods and Research*, *20* (3), 367–397.

Algina, J. (2000). Intraclass correlation – 3 level model (Message on Internet Discussion List, December 7, 2000). Multilevel Discussion List, archived at listserv@jiscmail.ac.uk.

Alker, H. R. (1969). A typology of fallacies. In M. Dogan & S. Rokkan (Eds.), *Quantitative ecological analysis in the social sciences*. Cambridge, MA: MIT Press.

Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: Sage.

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage.

Ansari, A., & Jedidi, K. (2000). Bayesian factor analysis for multilevel binary observations. *Psychometrika*, *65*, 475–496.

Anscombe, F. J. (1973). Graphs in statistical analysis. *American Statistician*, *27*, 17–21.

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Arbuckle, J. L. (2007). *Amos 16.0 user's guide*. Chicago: SPSS, Inc.

Arnold, C. L. (1992). An introduction to hierarchical linear models. *Measurement and Evaluation in Counseling and Development*, *25*, 58–90.

Asparouhov, T., & Muthén, B. (2007). Computationally efficient estimation of multilevel high-dimensional latent variable models. *Proceedings of the Joint Statistical Meeting*, August 2007, Salt Lake City, Utah. Accessed May 2009 at: http://www.statmodel.com.

Barber, J. S., Murphy, S., Axinn, W. G., & Maples, J. (2000) Discrete-time multi-level hazard analysis. *Sociological Methodology*, *30*, 201–235.

Barbosa, M. F., & Goldstein, H. (2000). Discrete response multilevel models for repeated measures: An application to voting intentions data. *Quality and Quantity*, *34*, 323–330.

Barnett, V. (1999). *Comparative statistical inference*. New York: Wiley.

Bauer, D. J., & Cai, L. (2009). Consequences of unmodeled nonlinear effects in multilevel models. *Journal of Educational and Behavioral Statistics*, *34*, 97–114.

Bauer, D. J., & Curran, P. J. (2005). Probing interactions in fixed and multilevel regression: Inferential and graphical techniques. *Multivariate Behavioral Research*, *40*, 373–400.

Bechger, T. M., van Schooten, E., de Glopper, C., & Hox, J. J. (1998). The validity of international surveys of reading literacy: The case of the IEA reading literacy study. *Studies in Educational Evaluation*, *24* (2), 99–125.

Beck, N., & Katz, J. N. (1997). *The analysis of binary time-series-cross-section data and/or the democratic peace*. Paper, Annual Meeting of the Political Methodology Group, Columbus, Ohio, July, 1997.

Becker, B. J. (1994). Combining significance levels. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 215–230). New York: Russel Sage Foundation.

Becker, B. J. (2007). Multivariate meta-analysis: Contributions by Ingram Olkin. *Statistical Science*, *22*, 401–406.

Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin*, *107*, 238–246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness-of-fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588–606.

Bentler, P. M., & Liang, J. (2001). General two level mean and covariance structure models: Maximum likelihood via EM-type algorithms. In J. Blasius, J. Hox, E. de Leeuw, & P. Schmidt (Eds.), *Social science methodology in the new millennium. Proceedings of the fifth international conference on logic and methodology*. Opladen, FRG: Leske + Budrich.

Berkey, C. S., Hoaglin, D. C., Antczak-Bouckoms, A., Mosteller, F., & Colditz, G. A. (1998). Meta-analysis of multiple outcomes by regression with random effects. *Statistics in Medicine*, *17*, 2537–2550.

Berkhof, J., & Snijders, T. A. B. (2001). Variance component testing in multilevel models. *Journal of Educational and Behavioral Statistics*, *26*, 133–152.

Biggerstaff, B. J., Tweedy, R. L., & Mengersen, K. L. (1994). Passive smoking in the workplace: Classical and Bayesian meta-analyses. *International Archives of Occupational and Environmental Health*, *66*, 269–277.

Blalock, H. M. (1990). Auxiliary measurement theories revisited. In J. J. Hox & J. De Jong-Gierveld (Eds.), *Operationalization and research strategy*. Amsterdam: Swets & Zeitlinger.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bollen, K. A., & Barb, K. H. (1981). Pearson's *r* and coarsely categorized measures. *American Sociological Review*, *46*, 232–239.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models*. New York: Wiley.

Bollen, K. A., & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods and Research*, *21*, 205–229.

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics*, *27*, 335–340.

Booth, J. G., & Sarkar, S. (1998). Monte Carlo approximation of bootstrap variances. *American Statistician*, *52*, 354.

Bosker, R. J., Snijders, T. A. B., & Guldemond, H. (2003). *User's manual PINT*. Program and manual available at: http://www.gamma.rug.nl/.

Boyd, L. H., & Iversen, G. R. (1979). *Contextual analysis: Concepts and statistical techniques*. Belmont, CA: Wadsworth.

Breslow, N. E., & Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika*, *82*, 81–91.

Brockwell, S. E., & Gordon, I. R. (2001). A comparison of statistical methods for meta-analysis. *Statistics in Medicine*, *20*, 825–840.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research*, *21*, 230–258.

Browne, W. J. (1998). *Applying MCMC methods to multilevel models*. Bath, UK: University of Bath.

Browne, W. J. (2005). *MCMC estimation in MLwiN, Version 2*. Bristol, UK: University of Bristol, Centre for Multilevel Modelling.

Browne, W. J., & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics*, *15*, 391–420.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, *101*, 147–158.

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

Burchinal, M., & Appelbaum, M. I. (1991). Estimating individual developmental functions: Methods and their assumptions. *Child Development*, *62*, 23–43.

Burton, P., Gurrin, L., & Sly, P. (1998). Extending the simple regression model to account for correlated responses: An introduction to generalized estimating equations and multi-level mixed modeling. *Statistics in Medicine*, *17*, 1261–1291.

Busing, F. (1993). *Distribution characteristics of variance estimates in two-level models*. Unpublished manuscript. Leiden: Department of Psychometrics and Research Methodology, Leiden University.

Camstra, A., & Boomsma, A. (1992). Cross-validation in regression and covariance structure analysis: An overview. *Sociological Methods and Research*, *21*, 89–115.

Capaldi, D. M., Crosby, L., & Stoolmiller, M. (1996). Predicting the timing of first sexual intercourse for at-risk adolescent males. *Child Development*, *67*, 344–359.

Carlin, B. P., & Louis, T. A. (1996). *Bayes and empirical Bayes methods for data analysis*. London: Chapman & Hall.

Carpenter, J., & Bithell, J. (2000). Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, *19*, 1141–1164.

Carpenter, J., Goldstein, H., & Rasbash, J. (2003). A novel bootstrap procedure for assessing the relationship between class size and achievement. *Applied Statistics*, *52*, 431–442.

Casella, G., & George, E. (1992). Explaining the Gibbs sampler. *American Statistician*, *46*, 167–174.

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology*, *83* (2), 234–246.

Cheong, Y. F., Fotiu, R. P., & Raudenbush, S. W. (2001). Efficiency and robustness of alternative estimators for two- and three-level models: The case of NAEP. *Journal of Educational and Behavioral Statistics*, *26*, 411–429.

Cheung, M. W. L. (2008). A model for integrating fixed-, random-, and mixed-effects meta-analysis into structural equation modeling. *Psychological Methods*, *13*, 182–202.

Chou, C.-P., & Bentler, P. M. (1995). Estimates and tests in structural equation modeling. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 37–55). Newbury Park, CA: Sage.

Chou, C.-P., Bentler, P. M., & Pentz, M. A. (2000). Comparisons of two statistical approaches to study growth curves: The multilevel model and latent curve analysis. *Structural Equation Modeling*, *5* (3), 247–267.

Cnaan, A., Laird, N. M., & Slasor, P. (1997). Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in Medicine*, *16*, 2349–2380.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin*, *70*, 426–443.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112* (1), 155–159.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cohen, M. P. (1998). Determining sample sizes for surveys with data analyzed by hierarchical linear models. *Journal of Official Statistics*, *14*, 267–275.

Cooper, H., Hedges, L. V., & Valentine, J. (Eds.). (2009). *The handbook of research synthesis and meta-analysis*. New York: Russel Sage Foundation.

Cornell, J., & Mulrow, C. (1999). Meta-analysis. In H. J. Ader & G. J. Mellenbergh (Eds.), *Research methodology in the social, behavioral, and life sciences*. London: Sage.

Cronbach, L. J. (1976). *Research in classrooms and schools: Formulation of questions, designs and analysis*. Occasional Paper, Stanford Evaluation Consortium.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measures*. New York: Wiley.

Cronbach, L. J., & Webb, N. (1979). Between class and within class effects in a reported aptitude × treatment interaction: A reanalysis of a study by G. L. Anderson. *Journal of Educational Psychology*, *67*, 717–724.

Croon, M. A., & van Veldhoven, M. J. P. M. (2007). Predicting group-level outcome variables from variables measured at the individual level: A latent variable multilevel model. *Psychological Methods*, *12*, 45–57.

Cudeck, R., & Klebe, K. J. (2002). Multiphase mixed-effects models for repeated measures data. *Psychological Methods*, *7*, 41–63.

Curran, P. J. (1997). *Supporting documentation for comparing three modern approaches to longitudinal data analysis: An examination of a single developmental sample*. Retrieved, June 2008 from http://www.unc.edu/~curran/srcd.html.

Curran, P. J. (2003). Have multilevel models been structural models all along? *Multivariate Behavioral Research*, *38*, 529–569.

Curran, P. J., & Bauer, D. J. (2007). Building path diagrams for multilevel models. *Psychological Methods*, *12*, 283–297.

Davidson, R., & MacKinnon, J. G. (1993). *Estimation and inference in econometrics*. New York: Oxford University Press.

Davis, P., & Scott, A. (1995). The effect of interviewer variance on domain comparisons. *Survey Methodology*, *21*, 2, 99–106.

De Leeuw, E. D. (1992). *Data quality in mail, telephone, and face-to-face surveys*. Amsterdam: TT-Publikaties.

De Leeuw, J. (2005). Dropouts in longitudinal data. In B. Everitt & D. Howell (eds), *Encyclopedia of statistics in behavioral science* (pp. 315–318). New York: Wiley.

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

Delucchi, K., & Bostrom, A. (1999). Small sample longitudinal clinical trials with missing data: A comparison of analytic methods. *Psychological Methods*, *4* (2), 158–172.

Diaz, R. E. (2007). Comparison of PQL and Laplace 6 estimates of hierarchical linear models when comparing groups of small incident rates in cluster randomised trials. *Computational Statistics and Data Analysis*, *51*, 2871–2888.

DiPrete, T. A., & Forristal, J. D. (1994). Multilevel models: Methods and substance. *Annual Review of Sociology*, *20*, 331–357.

DiPrete, T. A., & Grusky, D. B. (1990). The multilevel analysis of trends with repeated cross-sectional data. In C. C. Clogg (Ed.), *Sociological methodology*. London: Blackwell.

Dolan, C. V. (1994). Factor analysis with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology*, *47*, 309–326.

Dronkers, J., & Hox, J. J. (2006). The importance of common family background for the similarity of divorce risks of siblings: A multilevel event history analysis. In F. J. Yammarino & F. Dansereau (Eds.), *Multilevel issues in social systems* (pp. 217–238). Amsterdam: Elsevier.

Du Toit, M., & du Toit, S. (2001). *Interactive LISREL: User's guide*. Chicago: Scientific Software, Inc.

DuMouchel, W. H. (1989). Bayesian meta-analysis. In D. Berry (Ed.), *Statistical methodology in pharmaceutical sciences* (pp. 509–529). New York: Marcel Dekker.

DuMouchel, W. H. (1994). *Hierarchical Bayesian linear models for meta-analysis*. Unpublished report, Research Triangle Park, NC: National Institute of Statistical Sciences.

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2006). *An introduction to latent variable growth curve modeling*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Philadelphia: Society for Industrial and Applied Mathematics.

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

Eliason, S. R. (1993). *Maximum likelihood estimation*. Newbury Park, CA: Sage.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*, 121–138.

Engels, E. A., Schmidt, C. H., Terrin, N., Olkin, I., & Lau, J. (2000). Heterogeneity and statistical significance in meta-analysis: An empirical study of 125 meta-analyses. *Statistics in Medicine*, *19* (13), 1707–1728.

Erbring, L., & Young, A. A. (1979). Contextual effects as endogenous feedback. *Sociological Methods and Research*, *7*, 396–430.

Evans, M., Hastings, N., & Peacock, B. (2000). *Statistical distributions*. New York: Wiley.

Fears, T. R., Benichou, J., & Gail, M. H. (1996). A reminder of the fallibility of the Wald statistic. *American Statistician*, *50* (3), 226–227.

Field, A. (2009). *Discovering statistics using SPSS*. London: Sage.

Fielding, A. (2003). Ordered category responses and random effects in multilevel and other complex structures. In S. P. Reise & N. Duan (Eds.), *Multilevel modeling: Methodological advances, issues, and applications* (pp. 181–208). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Fielding, A. (2004). Scaling for residual variance components of ordered category responses in generalised linear mixed multilevel models. *Quality and Quantity*, *38*, 425–433.

Fotiu, R. P. (1989). *A comparison of the EM and data augmentation algorithms on simulated small sample hierarchical data from research on education*. Unpublished doctoral dissertation, Michigan State University, East Lansing.

Galtung, J. (1969). *Theory and methods of social research*. New York: Columbia University Press.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/ hierarchical models*. New York: Cambridge University Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–511.

Gerbing, D. W., & Anderson, J. C. (1992). Monte Carlo evaluations of goodness-of-fit indices for structural equation models. *Sociological Methods and Research*, *21*, 132–161.

Gibbons, R. D., & Bock, R. D. (1987). Trend in correlated proportions. *Psychometrika*, *52*, 113–124.

Gill, J. (2000). *Generalized linear models*. Thousand Oaks, CA: Sage.

Gleser, L. J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 339–355). New York: Russell Sage Foundation.

Glass, G. V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher*, *10*, 3–8.

Goldstein, H. (1987). *Multilevel models in educational and social research*. London: Griffin.

Goldstein, H. (1991) Non-linear multilevel models, with an application to discrete response data. *Biometrika*, *78*, 45–51.

Goldstein, H. (1994). Multilevel cross-classified models. *Sociological Methods and Research*, *22*, 364–376.

Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). London: Edward Arnold.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Edward Arnold.

Goldstein, H., & Browne, W. (2001). Multilevel factor analysis modeling using Markov chain Monte Carlo (MCMC) estimation. In G. A. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure models*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Goldstein, H., & Healy, M. J. R. (1995). The graphical representation of a collection of means. *Journal of the Royal Statistical Society, A*, *158*, 175–177.

Goldstein, H., Healy, M. J. R., & Rasbash, J. (1994). Multilevel time series models with applications to repeated measures data. *Statistics in Medicine*, *13*, 1643–1656.

Goldstein, H., & Rasbash, J. (1996). Improved approximations to multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A*, *159*, 505–513.

Goldstein, H., & Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society, A*, *159*, 505–513.

Goldstein, H., Steele, F., Rasbash, J., & Charlton, C. (2007). *REALCOM: Methodology for realistically complex multilevel modeling*. Bristol, UK: Centre for multilevel modeling, University of Bristol. Accessed May 2008 at: www.cmm.bristol.ac.uk.

Goldstein, H., Yang, M., Omar, R., Turner, R., & Thompson, S. (2000). Meta-analysis using multilevel models with an application to the study of class sizes. *Applied Statistics*, *49*, 399–412.

Good, P. I. (1999). *Resampling methods: A practical guide to data analysis*. Boston/Berlin: Birkhäuser.

Gray, B. R. (2005). Selecting a distributional assumption for modeling relative densities of benthic macroinvertebrates. *Ecological Modelling*, *185*, 1–12.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavior Research*, *26*, 499–510.

Greene, W. H. (1997). *Econometric analysis*. Upper Saddle River, NJ: Prentice Hall.

Grilli, L. (2005). The random effects proportional hazards model with grouped survival data: A comparison between the grouped continuous and continuation ratio versions. *Journal of the Royal Statistical Society, A*, *168*, 83–94.

Groves, R. M. (1989). *Survey cost and survey error*. New York: Wiley.

Gulliford, M. C., Ukoumunne, O. C., & Chinn, S. (1999). Components of variance and intraclass correlations for the design of community-based surveys and intervention studies. *American Journal of Epidemiology*, *149*, 876–883.

Gustafsson, J.-E., & Stahl, P. E. (1999). *Streams user's guide, Vs. 2.0*. Mölndal, Sweden: MultivariateWare.

Hartford, A., & Davidian, M. (2000). Consequences of misspecifying assumptions in nonlinear mixed effects models. *Computational Statistics and Data Analysis*, *34*, 139–164.

Harwell, M. (1997). An empirical study of Hedges' homogeneity test. *Psychological Methods*, *2* (2), 219–231.

Haughton, D. M. A., Oud, J. H. L., & Jansen, R. A. R. G. (1997). Information and other criteria in structural equation model selection. *Communications in Statistics: Simulation and Computation*, *26*, 1477–1516.

Hays, W. L. (1994). *Statistics*. New York: Harcourt Brace College Publishers.

Heck, R. H., & Thomas, S. L. (2009). *An introduction to multilevel modeling techniques*. New York: Routledge.

Hedeker, D. (2008). Multilevel models for ordinal and nominal variables. In J. de Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 239–276). New York: Springer.

Hedeker, D., & Gibbons, R. D. (1994). A random effects ordinal regression model for multilevel analysis. *Biometrics*, *50*, 933–944.

Hedeker, D., & Gibbons, R. D. (1996a). MIXOR: A computer program for mixed effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine*, *49*, 157–176.

Hedeker, D., & Gibbons, R. D. (1996b). MIXREG: A computer program for mixed effects regression analysis with autocorrelated errors. *Computer Methods and Programs in Biomedicine*, *49*, 229–252.

Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, *2* (1), 64–78.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. New York: Wiley.

Hedeker, D., Gibbons, R. D., du Toit, M., & Cheng, Y. (2008). *SuperMix: Mixed effects models*. Lincolnwood, IL: Scientific Software International.

Hedeker, D., & Mermelstein, R. J. (1998). A multilevel thresholds of change model for analysis of stages of change data. *Multivariate Behavioral Research*, *33*, 427–455.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random effects models in meta-analysis. *Psychological Methods*, *3*, 486–504.

Higgins, J. P. T., Whitehead, A., Turner, R. M., Omar, R. Z., & Thompson, S. G. (2001). Meta-analysis of continuous outcome data from individual patients. *Statistics in Medicine*, *20*, 2219–2241.

Hill, P. W., & Goldstein, H. (1998). Multilevel modeling of educational data with cross-classification and missing identification for units. *Journal of Educational and Behavioral Statistics*, *23* (2), 117–128.

Hoeksma, J. B., & Knol, D. L. (2001). Testing predictive developmental hypotheses. *Multivariate Behavior Research*, *36*, 227–248.

Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: implications for research in organizations. *Journal of Management*, *24* (5), 623–641.

Hoijtink, H. (2000). Posterior inference in the random intercept model based on samples obtained with Markov chain Monte Carlo methods. *Computational Statistics*, *15*, 315–336.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65–70.

Hox, J. J. (1993). Factor analysis of multilevel data: Gauging the Muthén model. In J. H. L. Oud & R. A. W. van Blokland-Vogelesang (Eds.), *Advances in longitudinal and multivariate analysis in the behavioral sciences* (pp. 141–156). Nijmegen: ITS.

Hox, J. J. (1995). *Applied multilevel analysis* (2nd ed.). Amsterdam: TT-Publikaties. Available at: http://www.geocities.com/joophox.

Hox, J. J. (1998). Multilevel modeling: When and why. In I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Classification, data analysis, and data highways* (pp. 147–154). New York: Springer Verlag.

Hox, J. J. (2001). *Estimating power in multilevel models: A general approach*. Paper presented at the 3rd Amsterdam multilevel conference, Amsterdam, April 9–10, 2001.

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hox, J. J., & Bechger, T. M. (1998). An introduction to structural equation modeling. *Family Science Review*, *11*, 354–373.

Hox, J. J., & de Leeuw, E. D. (1994). A comparison of nonresponse in mail, telephone, and face-to-face surveys. Applying multilevel modeling to meta-analysis. *Quality and Quantity*, *28*, 329–344.

Hox, J. J., & de Leeuw, E. D. (2003). Multilevel models for meta-analysis. In N. Duan & S. Reise (Eds.), *Multilevel modeling: Methodological advances, issues and applications*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hox, J. J., de Leeuw, E. D., & Kreft, G. G. (1991). The effect of interviewer and respondent characteristics on the quality of survey data: A multilevel model. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys*. New York: Wiley.

Hox, J. J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling*, *8*, 157–174.

Hox, J. J., & Maas, C. J. M. (2004). Multilevel structural equation models: The limited information approach and the multivariate multilevel approach. In K. van Montfoort, A. Satorra, & H. Oud (Eds.), *Recent developments in structural equation models* (pp. 135–149). Amsterdam: Kluwer.

Hox, J. J., & Maas, C. G. M. (2006). Multilevel models for multimethod measures. In M. Eid & E. Diener (Eds), *Multimethod measurement in psychology* (pp. 269–282). Washington, DC: American Psychological Association.

Hox, J. J., Maas, C. G. M., & Brinkhuis, M. J. S. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica* (in press).

Hu, F. B., Goldberg, J., Hedeker, D., Flay, B. R., & Pentz, M. A. (1998). Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. *American Journal of Epidemiology*, *147*, 694–703.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under non-standard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (pp. 221–233). Berkeley, CA: University of California Press.

Huedo-Medina, T. B., Sánchez-Meca, F., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I2 index? *Psychological Methods*, *11*, 193–206.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis* (2nd ed.). Newbury Park, CA: Sage.

Hüttner, H. J. M., & van den Eeden, P. (1993). *The multilevel design. A guide with an annotated bibliography, 1980–1993*. London: Greenwood Press.

Jaccard, J., Turrisi, R., & Wan, C. K. (1990). *Interaction effects in multiple regression*. Newbury Park, CA: Sage.

Jang, W., & Lim, J. (2009). A numerical study of PQL estimation biases in generalized linear mixed models under heterogeneity of random effects. *Communications in Statistics – Simulation and Computation*, *38*, 692–702.

Jedidi, K., & Ansari, A. (2001). Bayesian structural equation models for multilevel data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 129–157). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Johnson, D. R., & Creech, J. C. (1983). Ordinal measures in multiple indicator models: A simulation study of categorization error. *American Sociological Review*, *48*, 398–407.

Jöreskog, K. G., & Sörbom, D. (1989). *Lisrel 7: A guide to the program and applications*. Chicago: SPSS Inc.

Jöreskog, K. G., & Yang, F. (1996). Non-linear structural equation models: The Kenny-Judd model with interactions. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Kalaian, S. A., & Kasim, R. M. (2008). Multilevel methods for meta-analysis. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 315–344). Charlotte, NC: Information Age Publishing, Inc.

Kalaian, H. A., & Raudenbush, S. W. (1996). A multivariate mixed linear model for meta-analysis. *Psychological Methods*, *1*, 227–235.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*, 79–93.

Kaplan, D. (1995). Statistical power in SEM. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 100–117). Newbury Park, CA: Sage.

Kauermann, G., & Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, *96*, 1387–1396.

Kef, S., Habekothé, H. T., & Hox, J. J. (2000). Social networks of blind and visually impaired adolescents: Structure and effect on well-being. *Social Networks*, *22*, 73–91.

Kendall, M. G. (1959). Hiawatha designs an experiment. *American Statistician*, *13*, 23–24.

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. New York: Guilford Press.

Kieseppä, I. A. (2003). AIC and large samples. *Philosophy of Science*, *70*, 1265–1276.

King, G. (1997). *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data*. Princeton, NJ: Princeton University Press.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences*. Belmont, CA: Brooks/Cole.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Kish, L. (1987). *Statistical design for research*. New York: Wiley.

Klein, K. J., & Kozlowski, S. W. J. (2000). From micro to meso: Critical steps in conceptualizing and conducting multilevel research. *Organizational Research Methods*, *3*, 211–236.

Kline, R. B. (2004). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.

Kreft, I. G. G. (1996). *Are multilevel techniques necessary? An overview, including simulation studies*. Unpublished Report, California State University, Los Angeles. Available at: http://www.eric.ed.gov.

Kreft, I. G. G., & de Leeuw, E. D. (1987). The see-saw effect: A multilevel problem? A reanalysis of some findings of Hox and de Leeuw. *Quality and Quantity*, *22*, 127–137.

Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Newbury Park, CA: Sage.

Kreft, I. G. G., de Leeuw, J., & Aiken, L. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, *30*, 1–22.

Krüger, M. (1994). *Sekseverschillen in schoolleiderschap* [Gender differences in school leadership]. Alphen a/d Rijn: Samson.

LaHuis, D. M., & Ferguson, M. W. (2009). The accuracy of statistical tests for variance components is multilevel random coefficient modeling. *Organizational Research Methods*, *12*, 418–435.

Langford, I., & Lewis, T. (1998). Outliers in multilevel data. *Journal of the Royal Statistical Society, Series A*, *161*, 121–160.

Lazarsfeld, P. F., & Menzel, H. (1961). On the relation between individual and collective properties. In A. Etzioni (Ed.), *Complex organizations: A sociological reader*. New York: Holt, Rhinehart & Winston.

Lee A. H., Wang, K., Scott, J. A., Yau, K. K. W., & McLachlan, G. J. (2006). Multi-level zero-inflated Poisson regression modelling of correlated count data with excess zeros. *Statistical Methods in Medical Research*, *15*, 47–61.

Leppik, I. E., Dreifuss, F. E., Porter, R., Bowman, T., Santilli, N., Jacobs, M., et al. (1987). A controlled study of progabide in partial seizures: Methodology and results. *Neurology*, *37*, 963–968.

Lesaffre, E., & Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random-effects model: An example. *Applied Statistics*, *50*, 325–335.

Leyland, A. H. (2004) *A review of multilevel modelling in SPSS*. Retrieved September 2008 from http://www.cmm.bristol.ac.uk/learning-training/multilevel-m-software

Liang, K., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 45–51.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

Lillard, L. A., & Panis, C. W. A. (2000). *aML. Multilevel multiprocess statistical software, release 1*. Los Angeles, CA: EconWare.

Lindley, D. V., & Novick, M. R. (1981). The role of exchangeability in inference. *Annals of Statistics*, *9*, 45–58.

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, *34*, 1–41.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS system for mixed models*. Cary; NC: SAS Institute, Inc.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated measures studies. *Journal of the American Statistical Association*, *90*, 1112–1121.

Little, R. J. A., & Rubin, D. B. (1989). The treatment of missing data in multivariate analysis. *Sociological Methods and Research*, *18*, 292–326.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York: Wiley.

Little, T. D., Schnabel, K. U., & Baumert J. (Eds.). (2000). *Modeling longitudinal and multiple-group* *data. Practical issues, applied approaches, and specific examples*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Loehlin, J. C. (2004). *Latent variable models: An introduction to factor, path, and structural analysis*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, *54*, 217–224.

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.

Longford, N. T. (1993). *Random coefficient models*. Oxford, UK: Clarendon Press.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Maas, C. J. M., & Hox, J. J. (2004a). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, *58*, 127–137.

Maas, C. J. M., & Hox, J. J. (2004b). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics and Data Analysis*, *46*, 427–440.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*, 85–91.

Maas, C. J. M., & Snijders, T. A. B. (2003). The multilevel approach to repeated measures with missing data. *Quality and Quantity*, *37*, 71–89.

Macaskill, P., Walter, S. D., & Irwig, L. (2001). A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine*, *20*, 641–654.

Manor, O., and Zucker, D. M. (2004), Small sample inference for the fixed effects in the mixed linear model. *Computational Statistics and Data Analysis*, *46*, 801–817.

Marcoulides, G. A., & Schumacker, R. E. (Eds.). (2001). *New developments and techniques in structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Maxwell, S. E. (1998). Longitudinal designs in randomized group comparisons: When will intermediate observations increase statistical power? *Psychological Methods*, *3*, 275–290.

McCoach, D. B., & Black, A. C. (2008). Evaluation of model fit and adequacy. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 245–272). Charlotte, NC: Information Age Publishing.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman & Hall.

McDonald, R. P. (1994). The bilevel reticular action model for path analysis with latent variables. *Sociological Methods and Research*, *22*, 399–413.

McKelvey, R., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, *4*, 103–120.

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction*. New York: Guilford Press.

McLelland, G. H., and Judd, C. M. (1993). Statistical difficulties in detecting interactions and moderator effects. *Psychological Bulletin*, *114*, 376–390.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, *10*, 259–284.

Mehta, P. D., & West, S. G. (2000). Putting the individual back into individual growth curves. *Psychological Methods*, *5*, 23–43.

Menard, S. (1995). *Applied logistic regression analysis*. Thousand Oaks, CA: Sage.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55*, 107–122.

Meyer, D. K. (1964). A Bayesian school superintendent. *American Educational Research Journal*, *1*, 219–228.

Miyazaki, Y., & Raudenbush, S. W. (2000). Tests for linkage of multiple cohorts in an accelerated longitudinal design. *Psychological Methods*, *5*, 44–53.

Moerbeek, M. (2000). *The design and analysis of multilevel intervention studies*. Maastricht, the Netherlands, University of Maastricht, PhD Thesis.

Moerbeek, M., van Breukelen, G. J. P., & Berger, M. (2000). Design issues for experiments in multilevel populations. *Journal of Educational and Behavioral Statistics*, *25*, 271–284.

Moerbeek, M., van Breukelen, G. J. P., & Berger, M. (2001). Optimal experimental design for multilevel logistic models. *The Statistician*, *50*, 17–30.

Moghimbeigi, A., Eshraghian, M. R., Mohammad, K., & McArdle, B. (2008). Multilevel zero-inflated negative binomial regression modeling for over-dispersed count data with extra zeros. *Journal of Applied Statistics*, *10*, 1193–1202.

Moineddin, R., Matheson, F. I., & Glazier, R. H. (2007). A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology*, *7*, 34.

Mok, M. (1995). *Sample size requirements for 2-level designs in educational research*. Unpublished manuscript. London: Multilevel Models Project, Institute of Education, University of London.

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping. A nonparametric approach to statistical inference*. Newbury Park, CA: Sage.

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression*. Reading, MA: Addison-Wesley.

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*, 557–585.

Muthén, B. (1990). *Means and covariance structure analysis of hierarchical data*. Los Angeles: UCLA Statistics series, #62.

Muthén, B. O. (1991a). Analysis of longitudinal data using latent variable models with varying parameters. In L. C. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change*. Washington, DC: American Psychological Association.

Muthén, B. O. (1991b). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, *28*, 338–354.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research*, *22*, 376–398.

Muthén, B. O. (1997). Latent growth modeling with longitudinal and multilevel data. In A. E. Raftery (Ed.), *Sociological methodology, 1997* (pp. 453–480). Boston: Blackwell.

Muthén, B., du Toit, S. H. C., & Spisic, D. (1997), *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Accessed May 2007 at: http://www.statmodel.com.

Muthén, B. O., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, *38*, 171–189.

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. In P. V. Marsden (Ed.), *Sociological methodology, 1995* (pp. 267–316). Oxford, UK: Blackwell.

Muthén, L. K., & Muthén, B. O. (1998–2007). *Mplus user's guide*. Los Angeles: Muthén & Muthén.

Nam, I.-S., Mengersen, K., & Garthwaite, P. (2003). Multivariate meta-analysis. *Statistics in Medicine*, *22*, 2309–2333.

NCSS (1995). *NCSS probability calculator*. Available from NCSS website: http://www.ncss.com/download.html.

Nevitt, J., & Hancock, G. R. (2001). Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Structural Equation Modeling*, *8* (3), 353–377.

Normand, S.-L. (1999). Tutorial in biostatistics. Meta-analysis: Formulating, evaluating, combining, and reporting. *Statistics in Medicine*, *18*, 321–359.

Norusis, M. J. (2005). *SPSS 14.0 advanced statistical procedures companion*. Upper Saddle River, NJ: Prentice Hall.

Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research*. New York: McGraw-Hill.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.

O'Brien, R. G., & Kaiser, M. K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin*, *97*, 316–333.

Olsson, U. (1979). On the robustness of factor analysis against crude classification of the observations. *Multivariate Behavioral Research*, *14*, 485–500.

O'Muircheartaigh, C., & Campanelli, P. (1999). A multilevel exploration of the role of interviewers in survey non-response. *Journal of the Royal Statistical Society, Series A*, *162*, 437–446.

Paccagnella, O. (2006). Centering or not centering in multilevel models? *Evaluation Review*, *30*, 66–85.

Pan, H., & Goldstein, H. (1998). Multilevel repeated measures growth modeling using extended spline functions. *Statistics in Medicine*, *17*, 2755–2770.

Paterson, L. (1998). Multilevel multivariate regression: An illustration concerning school teachers' perception of their pupils. *Educational Research and Evaluation*, *4*, 126–142.

Pawitan, Y. (2000). A reminder of the fallibility of the Wald statistic: Likelihood explanation. *American Statistician*, *54* (1), 54–56.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction*. Fort Worth, TX: Harcourt.

Pendergast, J., Gange, S., Newton, M., Lindstrom, M., Palta, M., & Fisher, M. (1996). A survey of methods for analyzing clustered binary response data. *International Statistical Review*, *64* (1), 89–118.

Peugh, J. L., & Enders, C. K. (2005). Using the SPSS mixed procedure to fit cross-sectional and longitudinal multilevel models. *Educational and Psychological Measurement*, *65*, 714–741.

Pickery, J., & Loosveldt, G. (1998). The impact of respondent and interviewer characteristics on the number of 'no opinion' answers. *Quality and Quantity*, *32*, 31–45.

Pickery, J., Loosveldt, G., & Carton, A. (2001). The effects of interviewer and respondent characteristics on response behavior in panel-surveys: A multilevel approach. *Sociological Methods and Research*, *29* (4), 509–523.

Plewis, I. (1996). Statistical methods for understanding cognitive growth: A review, a synthesis and an application. *British Journal of Mathematical and Statistical Psychology*, *49*, 25–42.

Plewis, I. (2001). Explanatory models for relating growth processes. *Multivariate Behavior Research*, *36*, 207–225.

Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics*, *31*, 437–448.

Rabe-Hesketh, S., & Skrondal, A. (2008). *Multilevel and longitudinal modeling using stata* (2nd ed.). College Station, TX: Stata Press.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). *GLLAMM manual*. U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 160. Accessed May 2009 at: http://www.gllamm.org/docum.html.

Rabe-Hesketh, S., Skrondal, A., & Zheng, X. (2007). Multilevel structural equation modeling. In S.-Y. Lee (Ed.), *Handbook of latent variable and related models* (pp. 209–227). Amsterdam: Elsevier.

Raftery, A. E., & Lewis, S. M. (1992). How many iterations in the Gibbs sampler? In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4* (pp. 765–776). Oxford, UK: Oxford University Press.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., et al. (2000). *A user's guide to MLwiN*. London: Multilevel Models Project, University of London.

Rasbash, J., & Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *Journal of Educational and Behavioral Statistics*, *19* (4), 337–350.

Rasbash, J., Steele, F., Browne, W. J., & Goldstein, H. (2009). *A user's guide to MLwiN*. Bristol, UK: Centre for Multilevel Modelling, University of Bristol.

Raudenbush, S. W. (1993a). Hierarchical linear models as generalizations of certain common experimental designs. In L. Edwards (Ed.), *Applied analysis of variance in behavioral science* (pp. 459–496). New York: Marcel Dekker.

Raudenbush, S. W. (1993b). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics*, *18* (4), 321–349.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2* (2), 173–185.

Raudenbush, S. W. (2009). Analyzing effect sizes: Random effects. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 295–316). New York: Russell Sage Foundation.

Raudenbush, S., & Bhumirat, C. (1992). The distribution of resources for primary education and its consequences for educational achievement in Thailand. *International Journal of Educational Research*, *17*, 143–164.

Raudenbush, S. W., & Bryk, A. S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics*, *10*, 75–98.

Raudenbush, S. W., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, *59*, 1–17.

Raudenbush, S. W., & Bryk, A. S. (1988). Methodological advances in studying effects of schools and classrooms on student learning. *Review of Research on Education*, 15, 423–476.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., Bryk, A., Cheong, Y. F., & Congdon, R. (2004). *HLM 6. Hierarchical linear and nonlinear modeling*. Chicago: Scientific Software International.

Raudenbush, S. W., & Chan, W.-S. (1993). Application of a hierarchical linear model to the study of adolescent deviance in an overlapping cohort design. *Journal of Consulting and Clinical Psychology*, *61*, 941–951.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5* (2), 199–213.

Raudenbush, S. W., Rowan, B., & Kang, S. J. (1991) A multilevel, multivariate model for studying school climate with estimation via the EM algorithm and application to U.S. high-school data. *Journal of Educational Statistics*, *16* (4), 295–330.

Raudenbush, S. W., & Sampson, R. (1999a). Ecometrics: Toward a science of assessing ecological settings, with application to the systematic social observations of neighborhoods. *Sociological Methodology*, *29*, 1–41.

Raudenbush, S. W., & Sampson, R. (1999b). Assessing direct and indirect associations in multilevel designs with latent variables. *Sociological Methods and Research*, *28*, 123–153.

Raudenbush, S. W., & Willms, J. D. (Eds.). (1991). *Schools, classrooms, and pupils: International studies of schooling from a multilevel perspective*. New York: Academic Press.

Raudenbush, S. W., Yang, M.-L., & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics*, *9*, 141–157.

Reardon, S. F., Brennan, R., & Buka, S. I. (2002). Estimating multi-level discrete-time hazard models using cross-sectional data: Neighborhood effects on the onset of cigarette use. *Multivariate Behavioral Research*, *37*, 297–330.

Rijmen, F., Tuerlinckx, F., de Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, *8*, 185–205.

Riley, R. D., Thompson, J. R., & Abrams, K. R. (2008). An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. *Biostatistics*, *9*, 172–186.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, *15*, 351–357.

Rodriguez, G., & Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A*, *158*, 73–90.

Rodriguez, G., & Goldman, N. (2001). Improved estimation procedures for multilevel models with a binary response: A case study. *Journal of the Royal Statistical Society, Series A*, *164*, 339–355.

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Newbury Park, CA: Sage.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.

Sammel, M., Lin, X., & Ryan, L. (1999). Multivariate linear mixed models for multiple outcomes. *Statistics in Medicine*, *18*, 2479–2492.

Sampson, R., Raudenbush, S. W., & Earls, T. (1997). Neigborhoods and violent crime: A multilevel study of collective efficacy. *Science*, *227*, 918–924.

Sánchez-Meca, J., & Marín-Martínez, F. (1997). Homogeneity tests in meta analysis: A Monte Carlo comparison of statistical power and type I error. *Quality and Quantity*, *31*, 385–399.

Satorra, A. (1989). Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika*, *54*, 131–151.

Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, *50*, 83–90.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrika Bulletin*, *2*, 110–114.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, *78*, 719–727.

Schijf, B., & Dronkers, J. (1991). De invloed van richting en wijk op de loopbanen in de lagere scholen van de stad Groningen [The effect of denomination and neigborhood on education in basic schools in the city of Groningen in 1971]. In I. B. H. Abram, B. P. M. Creemers, & A. van der Ley (Eds.), *Onderwijsresearchdagen 1991: Curriculum*. Amsterdam: University of Amsterdam, SCO.

Schulze, R. (2008). *Meta-analysis, a comparison of approaches*. Göttingen, FRG: Hogrefe & Huber.

Schumacker, R. E., & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Schumacker, R. E., & Marcoulides, G. A. (Eds.). (1998). *Interaction and nonlinear effects in structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Schwarzer, R. (1989). *Meta-analysis programs. Program manual*. Berlijn: Institüt für Psychologie, Freie Universität Berlin.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York: Wiley.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlation: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.

Siddiqui, O., Hedeker, D., Flay, B. R., & Hu, F. B. (1996). Intraclass correlation estimates in a school-based smoking prevention study: Outcome and mediating variables, by gender and ethnicity. *American Journal of Epidemiology*, *144*, 425–433.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, *23*, 323–355.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford, UK: Oxford University Press.

Skinner, C. J., Holt, D., & Smith, T. M. F. (Eds.). (1989). *Analysis of complex surveys*. New York: Wiley.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.

Skrondal, A., & Rabe-Hesketh, S. (2007). Redundant overdisperion parameters in multilevel models for categorical responses. *Journal of Educational and Behavioral Statistics*, *32*, 419–430.

Smeeth, L., & Ng, E. S.-W. (2002). Intraclass correlations for cluster randomized trials in primary care: Data from the MRC trial of the assessment and management of older people in the community. *Controlled Clinical Trials*, *23*, 409–421.

Smith, A. F. M., & Gelfland, A. E. (1992). Bayesian statistics without tears: A sampling–resampling perspective. *American Statistician*, *46* (2), 84–88.

Smith, T. C., Spiegelhalter, D., & Thomas, A. (1995). Bayesian approaches to random-effects meta-analysis: A comparative study. *Statistics in Medicine*, *14*, 2685–2699.

Snijders, T. A. B. (1996). Analysis of longitudinal data using the hierarchical linear model. *Quality and Quantity*, *30*, 405–426.

Snijders, T. A. B., & Bosker, R. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics*, *18*, 237–259.

Snijders, T. A. B., & Bosker, R. (1994). Modeled variance in two-level models. *Sociological Methods and Research*, *22*, 342–363.

Snijders, T. A. B., & Bosker, R. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

Snijders, T. A. B., & Kenny, D. A. (1999). Multilevel models for relational data. *Personal Relationships*, *6*, 471–486.

Snijders, T. A. B., Spreen, M., & Zwaagstra, R. (1994). Networks of cocaine users in an urban area: The use of multilevel modelling for analysing personal networks. *Journal of Quantitative Anthropology*, *5*, 85–105.

Spiegelhalter, D. (1994). *BUGS: Bayesian inference using Gibbs sampling*. Cambridge, UK: MRC Biostatistics Unit. Available at: http://www.mrc-bsu.cam.ac.uk/bugs.

Spreen, M., & Zwaagstra, R. (1994). Personal network sampling, outdegree analysis and multilevel analysis: Introducing the network concept in studies of hidden populations. *International Sociology*, *9*, 475–491.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245–251.

Sterne, J. A. C., Becker, B. J., & Egger, M. (2005). The funnel plot. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds), *Publication bias in meta-analysis* (pp. 75–98). New York: Wiley.

Sterne, J. A. C., & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis* (pp. 99–110). New York: Wiley.

Stevens, J. (2009). *Applied multivariate statistics for the social sciences*. New York: Routledge.

Stinchcombe, A. L. (1968). *Constructing social theories*. New York: Harcourt.

Stine, R. (1989). An introduction to bootstrap methods. *Sociological Methods and Research*, *18*, 2–3, 243–291.

Stoel, R. D., Galindo, F., Dolan, C., & van den Wittenboer, G. (2006) On the likelihood ratio test when parameters are subject to boundary constraints. *Psychological Methods*, *11* (4), 439–455.

Stoel, R., & van den Wittenboer, G. (2001). Prediction of initial status and growth rate: Incorporating time in linear growth curve models. In J. Blasius, J. Hox, E. de Leeuw, & P. Schmidt (Eds.), *Social science methodology in the new millennium. Proceedings of the fifth international conference on logic and methodology*. Opladen, FRG: Leske + Budrich.

Sullivan, L. M., Dukes, K. A., & Losina, E. (1999). An introduction to hierarchical linear modeling. *Statistics in Medicine*, *18*, 855–888.

Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A., & Song, F. (2000). *Methods for meta-analysis in medical research*. New York: Wiley.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. New York: Pearson.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distribution by data augmentation [with discussion]. *Journal of the American Statistical Association*, *82*, 528–550.

Taris, T. W. (2000). *A primer in longitudinal data analysis*. Thousand Oaks, CA: Sage.

Tate, R. L., & Hokanson, J. E. (1993). Analyzing individual status and change with hierarchical linear models: Illustration with depression in college students. *Journal of Personality*, *61*, 181–206.

Tucker, C., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*, 1–10.

Turner, R. M., Omar, R. Z., Yang, M., Goldstein, H., & Thompson, S. G. (2000). A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine*, *19*, 3417–3432.

Van den Eeden, P., & Hüttner, H. J. M. (1982). Multi-level research. *Current Sociology*, *30* (3), 1–178.

Van der Leeden, R., & Busing, F. (1994). *First iteration versus IGLS/RIGLS estimates in two-level models: A Monte Carlo study with ML3*. Unpublished manuscript. Leiden: Department of Psychometrics and Research Methodology, Leiden University.

Van der Leeden, R., Busing, F., & Meijer, E. (1997). *Applications of bootstrap methods for two-level models*. Paper, Multilevel Conference, Amsterdam, April 1–2, 1997.

Van der Leeden, R., Meijer, E., & Busing, F., (2008). Resampling multilevel models. In J. de Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 401–433). New York: Springer.

Van Duijn, M. A. J., Snijders, T. A. B., & Zijlstra, B. J. H. (2004). $P_2$: A random effects model with covariates for directed graphs. *Statistica Neerlandica*, *58*, 234–254.

Van Duijn, M. A. J., van Busschbach, J. T., & Snijders, T. A. B. (1999). Multilevel analysis of personal networks as dependent variables. *Social Networks*, *21*, 187–209.

Van Houwelingen, H. C., Arends, L. R., & Stijnen, T. (2002). Advanced methods in meta-analysis: Multivariate approach and meta-regression. *Statistics in Medicine*, *21*, 589–624.

Van Peet, A. A. J. (1992). *De potentieeltheorie van intelligentie* [The potentiality theory of intelligence]. Amsterdam: University of Amsterdam, PhD Thesis.

Verbeke, G., & Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics and Data Analysis*, *23*, 541–556.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. Berlin: Springer.

Villar, J., Mackey, M. E., Carroli, G., & Donner, A. (2001). Meta-analyses in systematic reviews of randomized controlled trials in perinatal medicine: Comparison of fixed and random effects models. *Statistics in Medicine*, *20*, 3635–3647.

Wald, A. (1943): Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*, 426–482.

Walsh, J. E. (1947). Concerning the effect of intraclass correlation on certain significance tests. *Annals of Mathematical Statistics*, *18*, 88–96.

Wasserman, S., & Faust, K. (1994). *Social network analysis*. Cambridge, UK: Cambridge University Press.

West, B. T., Welch, K. B., & Gatecki, A. T. (2007). *Linear mixed models*. Boca Raton, FL: Chapman & Hall.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1–25.

Whitt, H. P. (1986). The sheaf coefficient: A simplified and expanded approach. *Social Science Research*, *15*, 175–189.

Willett, J. B. (1988). Questions and answers in the measurement of change. In E. Rothkopf (Ed.), *Review of research in education* (pp. 345–422). Washington, DC: AERA.

Willett, J. B. (1989). Some results on reliability for the longitudinal measurement of change: Implications for the design of studies of individual growth. *Educational and Psychological Measurement*, *49*, 587–602.

Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin*, *116*, 363–381.

Wolfinger, R. W. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika*, *80*, 791–705.

Wong, G. Y., & Mason, W. M. (1985) The hierarchical logistic regression model for multilevel analysis. Extensions of the hierarchical normal linear model for multilevel analysis. *Journal of the American Statistical Association*, *80*, 513–524.

Woodruff, S. I. (1997). Random-effects models for analyzing clustered data from a nutrition education intervention. *Evaluation Review*, *21*, 688–697.

Wright, D. B. (1997). Extra-binomial variation in multilevel logistic models with sparse structures. *British Journal of Mathematical and Statistical Psychology*, *50*, 21–29.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, *20*, 557–585.

Yuan, K.-H., & Hayashi, K. (2005), On Muthén's maximum likelihood for two-level covariance structure models. *Psychometrika*, *70*, 147–167.

Yung, Y.-F., & Chan, W. (1999). Statistical analyses using bootstrapping: Concepts and implementation. In R. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 82–108). Thousand Oaks, CA: Sage.

Zeger, S., Liang, K., & Albert, P. (1988). Models for longitudinal data: A generalised estimating equation approach. *Biometrics*, *44*, 1049–1060.

# Appendix A: Data and Stories

This appendix describes the data used for the examples in *Multilevel Analysis: Techniques and Applications, Second Edition*. Some of the examples are real data; other data sets have been simulated especially for their use in this book. The simulated data sets have been constructed following some hypothetical but plausible real-world scenario. This appendix describes the various data sets, giving either a reference to the study where they come from, or the 'story' that has been used as a template to generate the data.

Data are currently available on the Internet in SPSS system-file and portable file format, and in addition in the format in which they were analyzed for the book (e.g. HLM or MLwiN files). Most analyses in this book can be carried out by the majority of the available multilevel software. Obviously, there is a limit to the number of computer packages one can master. Most of the multilevel regression analyses in this book have been carried out in both HLM and MLwiN, and the multilevel SEM analyses have been carried out using LISREL and Mplus. System files and setups using these packages, where present, will also be made available on the Internet (currently http://www.joophox.net). I invite users of other multilevel software to use these data for their own learning or teaching. I appreciate receiving data sets and setups that have been transferred to other software systems, so I can make them also available to other users.

The format of the variables is chosen in such a way that writing the variables out in ASCII format results in a file where all variables are separated by at least one space. This file can be read into other programs using the *free format* option. The data sets are described in the order that they are introduced in the book.

## POPULARITY DATA

The popularity data in *popular2.** are simulated data for 2000 pupils in 100 schools. The purpose is to offer a very simple example for multilevel regression analysis. The main outcome variable is the *pupil popularity*, a popularity rating on a scale of 1–10 derived by a sociometric procedure. Typically, a sociometric procedure asks all pupils in a class to rate all the other pupils, and then assigns the average received popularity rating to each pupil. Because of the sociometric procedure, group effects as apparent from higher-level variance components are rather strong. There is a second outcome variable: pupil popularity as rated by their teacher, on a scale from 1 to 10. The explanatory variables are pupil gender (boy = 0, girl = 1), pupil extraversion (10-point scale), and teacher experience in years. The pupil popularity data are used as the main example in Chapter 2. They could also be used with both outcome variables as an example for the multilevel multivariate analysis in Chapter 10 (Chapter 10 uses the survey meta-analysis data for that purpose; a multivariate multilevel analysis of the popularity data is left as an exercise for the reader). These data are also used as the vehicle to compare the different estimation and testing procedures described in Chapter 3 and Chapter 13. The popularity data have been generated to be a 'nice' well-behaved data set: the sample sizes at both levels are sufficient, the residuals have a normal distribution, and the multilevel effects are strong.

## NURSES

The files *nurses*.* contains three-level data from a hypothetical study on stress in hospitals. The data are from nurses working in wards nested within hospitals. It is a cluster-randomized experiment. In each of 25 hospitals, four wards are selected and randomly assigned to an experimental and a control condition. In the experimental condition, a training program is offered to all nurses to cope with job-related stress. After the program is completed, a sample of about 10 nurses from each ward is given a test that measures job-related stress. Additional variables are: nurse age (years), nurse experience (years), nurse gender (0 = male, 1 = female), type of ward (0 = general care, 1 = special care), and hospital size (0 = small, 1 = medium, 2 = large). The data have been generated to illustrate three-level analysis with a random slope for the effect of *ExpCon*.

## GPA DATA

The GPA data are a longitudinal data set, where 200 college students have been followed for six consecutive semesters. The data are simulated. In this data set there are GPA measures taken on six consecutive occasions, with a job status variable (how many hours worked) for the same six occasions. There are two student-level explanatory variables: the gender (0 = male, 1 = female) and the high school GPA. These data are used in the longitudinal analyses in Chapter 5, and again in the latent curve analysis in Chapter 14. There is also a dichotomous student-level outcome variable, which indicates whether a student has been admitted to the university of their choice. Since not every student applies to a university, this variable has many missing values. The outcome variable 'admitted' is not used in any of the examples in this book.

These data come in several varieties. The basic data file is *gpa*. In this file, the six measurement occasions are represented by separate variables. Some software packages (e.g., Prelis) use this format. Other multilevel software packages (HLM, MLwiN, MixReg, SAS) require that the separate measurement occasions are different data records. The GPA data arranged in this ‘long’ data format are in the data file *gpalong*. A second data set based on the GPA data involves a process of panel attrition being simulated. Students were simulated to drop out, partly based on having a low GPA in the previous semester. This dropout process leads to data that are missing at random (MAR). A naive analysis of the incomplete data gives biased results. A sophisticated analysis using multilevel longitudinal modeling or SEM with the modern raw data likelihood (available in AMOS, Mplus and MX, and in recent versions of LISREL) should give unbiased results. Comparing analyses on the complete and the incomplete data sets gives an impression of the amount of bias. The incomplete data are in files *gpamiss* and *gpamislong*.

## THE CURRAN LONGITUDINAL DATA

The data in the SPSS file(s) *curran*.sav* are a data set constructed by Patrick Curran for a symposium *Comparing Three Modern Approaches to Longitudinal Data Analysis: An Examination of a Single Developmental Sample* conducted at the 1997 Biennial Meeting of the Society for Research in Child Development. In this symposium, several different approaches to longitudinal modeling (latent growth curves, multilevel analysis, and mixture modeling) were compared and contrasted by letting experts analyze a single shared data set. This data set, hereafter called the *curran data*, was compiled by Patrick Curran from a large longitudinal data set. Supporting documentation and the original data files are available on the Internet (http://www.unc.edu/~curran/srcd.html).

The data are a sample of 405 children who were within the first 2 years of entry to elementary school. The data consist of four repeated measures of both the child’s antisocial behavior and the child’s reading recognition skills. In addition, on the first measurement occasion, measures were collected of emotional support and cognitive stimulation provided by the mother. The data were collected using face-to-face interviews of both the child and the mother at 2-year intervals between 1986 and 1992.

## THAILAND EDUCATION DATA

The Thailand education data in file *thaieduc* are one of the example data sets that are included with the software HLM (also in the student version of HLM). They are discussed at length in the HLM user’s manual. They stem from a large survey of primary education in Thailand (Raudenbush & Bhumirat, 1992). The outcome variable is dichotomous, an indicator whether a pupil has ever repeated a class (0 = no, 1 = yes). The explanatory variables are pupil gender (0 = girl, 1 = boy), pupil pre-primary education (0 = no, 1 = yes), and the school's mean SES. The example in Chapter 6 of this book uses only pupil gender as explanatory variable. There are 8582 cases in the file *thaieduc*, but school mean SES is missing in some cases; there are 7516 pupils with complete data.

Note that these missing data have to be dealt with before the data are transported to a multilevel program. In the analysis in Chapter 6 they are simply removed using listwise deletion. However, the percentage of pupils with incomplete data is 12.4%, which is too large to be simply ignored in a real analysis.

## SURVEY RESPONSE META-ANALYSIS DATA

The survey response data used to analyze proportions in Chapter 6 are from a meta-analysis by Hox and de Leeuw (1994). The basic data file is *metaresp*. This file contains an identification variable for each study located in the meta-analysis. A mode identification indicates the data collection mode (face-to-face, telephone, mail). The main response variable is the proportion of sampled respondents who participate. Different studies report different types of response proportions: we have the completion rate (the proportion of participants from the total initial sample) and the response rate (the proportion of participants from the sample without ineligible respondents (moved, deceased, address nonexistent). Obviously, the response rate is usually higher than the completion rate. The explanatory variables are the year of publication and the (estimated) saliency of the survey's main topic. The file also contains the denominators for the completion rate and the response rate, if known. Since most studies report only one of the response figures, the variables 'comp' and 'resp' and the denominators have many missing values.

Some software (e.g., MLwiN) expects the *proportion* of 'successes' and the denominator on which it is based; other software (e.g., HLM) expects the *number* of 'successes' and the corresponding denominator. The file contains the proportion only; the number of successes must be computed from the proportion if the software needs that. The file *multresp* contains the same information, but now in a three-level format useful if the data are analyzed using the multivariate outcome, which is demonstrated in Chapter 11.

## STREET SAFETY DATA

A sample of 100 streets are selected, and on each street a random sample of 10 persons are asked how often they feel unsafe while walking that street. The question about feeling unsafe is asked using three answer categories: 1 = never, 2 = sometimes, 3 = often. Predictor variables are age and gender; street characteristics are an economic index (standardized $Z$-score) and a rating of the crowdedness of the street (7-point scale). File: *Safety*. Used in Chapter 7 on ordinal data.

## EPILEPSY DATA

The epilepsy data come from a study by Leppik et al. (1987). They have been analyzed by many authors, including Skrondal and Rabe-Hesketh (2004). The data come from a randomized controlled study on the effect of an anti-epileptic drug versus a placebo. It is a longitudinal design. For each patient the number of seizures was measured for a 2-week baseline. Next, patients were randomized to the drug or the placebo condition. For four consecutive visits the clinic collected counts of epileptic seizures in the 2 weeks before the visit. The data set contains the following variables: count of seizures, treatment indicator, visit number, dummy for visit #4, log of age, log of baseline count. All predictors are grand mean centered. The data come from the GLLAMM homepage at: www.gllamm.org/books, used in Chapter 7 on count data.

## FIRST SEX DATA

This is a data set from Singer and Willett's book on longitudinal data analysis (2003), from a study by Capaldi, Crosby, and Stoolmiller (1996). A sample of 180 middle-school boys were tracked from the 7th through the 12th grade, with the outcome measure being when they had sex for the first time. At the end, 54 boys (30%) were still virgins. These observations are censored. File *firstsex* is used as an example of (single-level) survival data in Chapter 8. There is one dichotomous predictor variable, which is whether there has been a parental transition (0 if the boy lived with his biological parents before the data collection began).

## SIBLING DIVORCE

This involves multilevel survival data analyzed by Dronkers and Hox (2006). The data are from the National Social Science Family Survey of Australia of 1989–1990. In this survey detailed information was collected, including the educational attainment of respondents, their social and economic background, such as parental education and occupational status of the father, parental family size and family form, and other relevant characteristics of 4513 men and women in Australia. The respondent also answered all these questions about his or her parents and siblings. The respondents gave information about at most three siblings, even if there were more siblings in the family. All sibling variables were coded in the same way as the respondents, and all data were combined in a file with respondents or siblings as the unit of analysis. In that new file, respondents and siblings from the same family had the same values for their parental characteristics, but had different values for their child characteristics. The data file contains only those respondents or siblings that were married or had been married, and gave no missing values. File: *sibdiv*.

## PUPCROSS DATA

This data file is used to demonstrate the cross-classified data with pupils nested within both primary and secondary schools. These are simulated data, where 1000 pupils attended 100 primary and subsequently 30 secondary schools. There is no complete nesting structure; the pupils are nested within the cross-classification of primary and secondary schools. The file *pupcross* contains the secondary school achievement score, which is the outcome variable, and the explanatory pupil-level variables, gender (0 = boy, 1 = girl) and SES. School-level explanatory variables are the denomination of the primary and the secondary school (0 = no, 1 = yes). These data are used for the example of a cross-classified analysis in Chapter 8.

## SOCIOMETRIC SCORES DATA

The sociometric data are simulated data, intended to demonstrate a data structure where the cross-classification is at the lowest level, with an added group structure because there are several groups. The story is that in small groups all members are asked to rate each other. Since the groups are of different sizes, the usual data file organized by case in *socscors* has many missing values. The data are rearranged in a data file called *soclong* for the multilevel analysis. In *soclong* each record is defined by the sender–receiver pairs, with explanatory variables age and sex defined separately for the sender and the receiver. The group variable ‘group size’ is added to this file.

## SCHOOL MANAGER DATA

The school manager data are from an educational research study (Krüger, 1994). In this study, male and female school managers from 98 schools were rated by 854 pupils. The data are in file *manager*. These data are used to demonstrate the use of multilevel regression modeling for measuring context characteristics (here, the school manager’s management style). The questions about the school manager are questions 5, 9, 12, 16, 21, and 25; in Chapter 10 of the book these are renumbered 1 . . . 6. These data are used only to demonstrate the multilevel psychometric analyses in Chapter 9. They can also be analyzed using one of the multilevel factor analysis procedures outlined in Chapter 12. The data set also contains the pupils' and school manager's gender (1 = female, 2 = male), which is not used in the example. The remaining questions in the data set are all about various aspects of the school environment; a full multilevel exploratory factor analysis is a useful approach to these data.

## SOCIAL SKILLS META-ANALYSIS DATA

The social skills meta-analysis data in file *meta20* contain the coded outcomes of 20 studies that investigate the effect of social skills training on social anxiety. All studies use an experimental group/control group design. Explanatory variables are the duration of the training in weeks, the reliability of the social anxiety measure used in each study (two values, taken from the official test manual), and the studies' sample size. The data are simulated.

## ASTHMA AND LRD META-ANALYSIS DATA

The asthma and LRD data are from Nam, Mengersen, and Garthwaite (2003). The data are from a set of studies that investigate the relationship between children's environmental exposure to smoking (ETS) and the child health outcomes asthma and lower respiratory disease (LRD). Available are the logged odds ratio (LOR) for asthma and LRD, and their standard errors. Study-level variables are the average age of subjects, publication year, smoking (0 = parents, 1 = other in household), and covariate adjustment used (0 = no, 1 = yes).

There are two effect sizes, the logged odds ratio for asthma and lower respiratory disease (LRD). Only a few studies report both. Datafile: *AstLrd*.

## ESTRONE DATA

The estrone data are 16 independent measurements of the estrone level of five post-menopausal women (Fears et al., 1996). The data file *estronex* contains the data in the usual format; the file *estrlong* contains the data in the format used for multilevel analysis. Although the data structure suggests a temporal order in the measurements, there is none. Before the analysis, the estrone levels are transformed by taking the natural logarithm of the measurements. The estrone data are used in Chapter 13 to illustrate the use of advanced estimation and testing methods on difficult data. The difficulty of the estrone data lies in the extremely small sample size and the small value of the variance components.

## GOOD89 DATA

The file *good89* (from Good, 1999, p. 89) contains the very small data set used to demonstrate the principles of bootstrapping in Chapter 13.

## FAMILY IQ DATA

The family IQ data are patterned to follow the results from a study of intelligence in large families (van Peet, 1992). They are the scores on six subscales from an intelligence test and are used in Chapter 14 to illustrate multilevel factor analysis. The file *FamilyIQ* contains the data from 275 children in 50 families. The data file contains the additional variables gender and parental IQs, which are not used in the analyses in this book. Datafile: *FamIQ*.

## GALO DATA

The GALO data in file *galo* are from an educational study by Schijf and Dronkers (1991). They are data from 1377 pupils within 58 schools. We have the following pupil-level variables: father's occupational status *focc*, father's education *feduc*, mother's education *meduc*, pupil sex *sex*, the result of GALO school achievement test *GALO*, and the teacher's advice about secondary education *advice*. At the school level we have only one variable: the school's denomination *denom*. Denomination is coded 1 = Protestant, 2 = Nondenominational, 3 = Catholic (categories based on optimal scaling). The data file *galo* contains both complete and incomplete cases, and an indicator variable that specifies whether a specific case in the data file is complete or not.

# Appendix B: Aggregating and Disaggregating

A common procedure in multilevel analysis is to aggregate individual-level variables to higher levels. In most cases, aggregation is used to attach to higher-level units (e.g., groups, classes, teachers) the mean value of a lower-level explanatory variable. However, other aggregation functions may also be useful. For instance, one may have the hypothesis that classes that are heterogeneous with respect to some variable differ from more homogeneous classes. In this case, the aggregated explanatory variable would be the group's standard deviation or the range of the individual variable. Another aggregated value that can be useful is the group size.

In SPSS, aggregation is handled by the procedure *aggregate*. This procedure produces a new file that contains the grouping variable and the (new) aggregated variables. In SPSS/Windows *aggregate* is available in the DATA menu. A simple syntax to aggregate the variable IQ in a file with grouping variable groupnr is as follows:

```
GET FILE 'indfile.sys'.
AGGREGATE OUTFILE='aggfile.sys'/BREAK=groupnr/
meaniq=MEAN(iq)/stdeviq=SD(iq).
```

Disaggregation means adding group-level variables to the individual data file. This creates a file where the group-level variables are repeated for all individuals in the same group. In SPSS, this can be accomplished by the procedure JOIN MATCH, using the so-called TABLE lookup. Before JOIN MATCH is used, the individual and the group file must both be sorted on the group identification variable. In SPSS/Windows JOIN MATCH is available in the DATA menu. For instance, if we want to read the aggregated mean IQ and IQ standard deviation to the individual file, we have the following setup:

```
JOIN MATCH FILE='indfile.sys'/ TABLE='aggfile.sys'/ BY groupnr/MAP.
```

The example below is a complete setup that uses aggregation and disaggregation to get group means and individual deviation scores for IQ:

```
GET FILE 'indfile.sys'.
SORT groupnr.
SAVE FILE 'indfile.sys'.
AGGREGATE OUTFILE='aggfile.sys'/PRESORTED/BREAK=groupnr/
```

```
meaniq=MEAN(iq)/stdeviq=SD(iq).
JOIN MATCH FILE=‘indfile.sys’/ TABLE=‘aggfile.sys’/ BY groupnr/MAP.
COMPUTE deviq=iq-meaniq.
SAVE FILE ‘indfile2.sys’.
```

In this setup I use the AGGREGATE subcommand PRESORTED to indicate that the file is already sorted on the BREAK variable groupnr, because this saves computing time. The subcommand MAP on the JOIN MATCH procedure creates a map of the new system file, indicating from which of the two old system files the variables are taken. In this kind of ‘cutting and pasting’ it is extremely important to check the output of both AGGREGATE and JOIN MATCH very carefully to make sure that the cases are indeed matched correctly.

It should be noted that the program HLM contains a built-in procedure for centering explanatory variables. The program MLwiN has a procedure to add group means to the individual data file, and commands to create centered and group-centered variables.

A particular form of disaggregation is when we have a file with repeated measures, with repeated measures represented by separate variables. Many programs need data where each measurement occasion is seen as a separate row of data, with time-invariant variables repeated in the new data file. The GPA data are a good example. To create the ‘long’ data file needed I used the following SPSS syntax:

```
GET FILE ‘d:\data\gpa.sav’.
WRITE OUTFILE ‘d:\data\gpalong.dat’ RECORDS=6/
student ‘0’ gpa1 job1 sex highgpa /
student ‘1’ gpa2 job2 sex highgpa /
student ‘2’ gpa3 job3 sex highgpa /
student ‘3’ gpa4 job4 sex highgpa /
student ‘4’ gpa5 job5 sex highgpa /
student ‘5’ gpa6 job6 sex highgpa.
EXECUTE.
DATA LIST FILE ‘d:\ data\gpalong.dat’ FREE /
student occasion gpa job sex highgpa.
SAVE OUTFILE ‘d:\ data\gpalong.sav’.
DESCRIPTIVES ALL.
```

This syntax first writes out the data in ASCII format, and then reads these back into SPSS using the DATA LIST command with a different structure. The final command DESCRIPTIVES is used to check if all variables have plausible values.

A complication arises if the original data file has missing values. These are often coded in SPSS as system missing values, which are written out as *spaces*. When the command DATA LIST 'filename' FREE / is used in SPSS, these are read over and after the first such occurrence all other variables have incorrect values. To prevent this, we need to insert a command that recodes all system missing values into a real code, and after creating the flat data file, records that contain such missing value codes must be removed. To create the 'long' data file needed from the incomplete data set gpamiss I used the following SPSS syntax:

```
GET FILE 'd:\data\gpamiss.sav'.
RECODE gpa1 to job6 (SYSMIS=9).
WRITE OUTFILE 'd:\joop\Lea\data\mislong.dat' RECORDS=6/
student '0' gpa1 job1 sex highgpa /
student '1' gpa2 job2 sex highgpa /
student '2' gpa3 job3 sex highgpa /
student '3' gpa4 job4 sex highgpa /
student '4' gpa5 job5 sex highgpa /
student '5' gpa6 job6 sex highgpa.
EXECUTE.
DATA LIST FILE 'd:\ data\mislong.dat' FREE /
student occasion gpa job sex highgpa.
COUNT out=gpa job (9).
SELECT IF (out=0).
SAVE OUTFILE 'd:\ data\mislong.sav'.
```

# Appendix C: Recoding Categorical Data

Including categorical data in a linear regression system has been discussed in detail by Cohen (1968), and more recently in Pedhazur (1997). There was great interest in these methods in the 1960s because there was a lack of software for analysis of variance designs, and this instigated much interest in coding categorical predictor variables for analysis using multiple regression methods. Interestingly, interest in these coding methods has returned, since multilevel regression and (multilevel) structural equation modeling are also regression models, and categorical variables need to be given special codes to include them in these models.

## DUMMY CODING

The simplest way to code a categorical variable is dummy coding, which assigns a '1' to indicate the presence in a specific category, and a '0' to indicate absence. Assume that we have a treatment variable with three categories, with 1 indicating no treatment, 2 indicating treatment A, and 3 indicating treatment B. For a categorical variable with three categories we need two dummy variables. One category is the reference category, which is coded '0' for both dummy variables. Since category 1 is the control group, it makes sense to designate this category as the reference category, and we get the coding in Table C.1.

*Table C.1* Dummy coding for two treatments and one control group

| Treatment | Dummy1 | Dummy2 |
|---|---|---|
| 1 = Control | 0 | 0 |
| 2 = Treatment A | 1 | 0 |
| 3 = Treatment B | 0 | 1 |

Since dummy1 refers to treatment A and dummy2 to treatment B, we can rename the dummies as TreatA and TreatB, which directly reflects their meaning. In a regression with only the two dummies, the intercept is the predicted outcome for the control group, and the regression coefficients for the dummies indicate how much treatment A or B adds to or subtracts from the control group mean. If we remove the intercept from the model, we can include dummy variables for all available categories.

## EFFECT CODING

A different way to code a categorical variable is effect coding. With effect coding, the effect codes take on values '–1', '0', and '1', with the reference category coded '–1' for all effect variables. This produces the coding in Table C.2.

*Table C.2* Effect coding for two treatments and one control group

| Treatment | Effect1 | Effect2 |
|---|---|---|
| 1 = Control | –1 | –1 |
| 2 = Treatment A | 1 | 0 |
| 3 = Treatment B | 0 | 1 |

With effect coding, the intercept represents the grand mean, and the regression coefficients of the effect variables reflect how much the indicated treatment adds to or subtracts from the grand mean. In analysis of variance the deviation of a cell from the grand mean is called the treatment effect, hence the name 'effect coding' for this system. When there is an unambiguous control group, effect coding is less useful than simple dummy coding. However, if there are three treatments in our example, and no real control group, effect coding is a useful way to analyze which treatment has the largest effect. In this case, leaving out the intercept and including all possible effect codes is very effective.

## CONTRAST CODING

Contrast coding is mainly used to code for a specific hypothesis. It is a powerful and flexible coding method for constructing and testing relationships between a categorical variable and a continuous variable. For example, assume that treatments A and B in our example are a vitamin supplement, with treatment B being a more powerful dose than treatment A. Reasonable hypotheses about this experiment are: (1) Does supplementing vitamins help at all? and (2) Does an increased dose have a larger effect? This can be coded in the contrast codes shown in Table C.3.

*Table C.3* Contrast coding for two treatments and one control group

| Treatment | Contrast1 | Contrast2 |
|---|---|---|
| 1 = Control | –1 | 0 |
| 2 = Treatment A | .5 | –.5 |
| 3 = Treatment B | .5 | .5 |

In this scheme, the first contrast compares the control group to both treatment groups, and the second contrast compares treatment A against treatment B. In general, contrast codes represent hypotheses by a set of contrast weights. Positive and negative weights are used to indicate which groups or categories are compared or contrasted. A usual requirement for contrast weights is that they sum to zero across all groups or categories. Sometimes contrast codes are constructed that are orthogonal, that is, mutually independent. This is the case if the sum of the products of the contrast weights equals zero. Independence of contrast codes is nice, but it is more important that the codes reflect the hypotheses of interest accurately.

## NOTES

In addition to the coding methods discussed above, there are other coding methods, such as difference coding and (reverse) Helmert coding. These are useful in special situations, and are discussed in the ANOVA literature because they are used for planned or post hoc testing. One coding system that is often used for trend analysis is orthogonal polynomials. These are discussed in Appendix D.

When a categorical variable is represented by a set of codes, one would generally treat these coded variables as a block, with an overall significance test rather than separate tests for each. This is especially the case with dummy or effect coding; with contrast codes that each represent a distinct hypothesis, testing the contrasts separately is common.

Interactions with coded categorical variables follow the same rules as interactions for continuous variables (see Aiken & West, 1991, for a discussion).

When using statistical software that allows categorical variables (often called ‘factors’) in regression models, it is important to know what the default coding method is for categorical variables in regression models, and which category is treated as the reference category. In many cases, the default option may not be optimal for the analysis problem at hand, and manually coding the preferred coding method is better.

# Appendix D: Constructing Orthogonal Polynomials

When a categorical variable represents increasing levels of a dosage, or successive measurement occasions, the trend is often studied using polynomial regression. Thus, if the measurement occasion is represented by the variable $t$, the polynomial regression model will have the form $Y = b_0 + b_1 t + b_2 t^2 + \ldots, + b_T t^T$. The graph of $Y$ against $t$ will be curved. Nevertheless, the polynomial equation is not strictly a nonlinear equation; mathematically it is linear. This makes polynomial regression a very flexible way to model complex curves, or to approximate a smooth curve (it is used for this purpose in discrete time survival analysis; see Chapter 8 for an example). The disadvantage of polynomial regression is that few processes under observation actually follow polynomial curves. This means that polynomial results can rarely be interpreted in theoretical terms.

A disadvantage of polynomials as simple powers of the time variable is that these predictors tend to be highly correlated. Therefore, *orthogonal polynomials* are used, which are transformed values that reflect the various degrees of the simple polynomial, but are uncorrelated.

When the levels of the categorical variable are evenly spaced, and there are equal numbers of subjects in the categories, constructing orthogonal polynomials is straightforward. For unequal spacing or unequal *n*s, adaptations exist (see Cohen & Cohen, 1983; Kirk, 1982). However, if the unbalance is not extreme, standard orthogonal polynomials will still work well, even though they are not really orthogonal anymore.

How to construct orthogonal polynomials using matrix procedures built into the major software packages like SPSS and SAS is explained in detail by Hedeker and Gibbons (2006). Their procedure requires knowledge of matrix algebra, but has the advantage that it can deal with uneven spacing of the time variable. For our example, we construct orthogonal polynomials for the longitudinal GPA data analyzed in the chapter on longitudinal models. In this example, we have GPA data for students in six consecutive semesters. Table D.1 below codes the measurement occasion as 0–5 and lists all possible polynomials.

Table D.2 shows the correlations between these four time measures.

The correlations between the simple polynomials are very high, and certain to produce collinearity problems when these polynomials are all used together. Since we have evenly spaced measurement occasions, we can use standard tables for orthogonal polynomials, for example as given by Pedhazur (1997). The four orthogonal polynomials are shown in Table D.3.

The correlations between these orthogonal polynomials are indeed all zero. Hedeker and Gibbons suggest dividing each of these polynomial variables by the

*Table D.1* Four simple polynomials

| T1 | T2 | T3 | T4 |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 1 |
| 2 | 4 | 8 | 16 |
| 3 | 9 | 27 | 81 |
| 4 | 16 | 64 | 256 |
| 5 | 25 | 125 | 625 |

*Table D.2* Correlations between four simple polynomials

| | t1 | t2 | t3 | t4 |
|---|---|---|---|---|
| t1 | 1.00 | .96 | .91 | .86 |
| t2 | .96 | 1.00 | .99 | .96 |
| t3 | .91 | .99 | 1.00 | .99 |
| t4 | .86 | .96 | .99 | 1.00 |
| t5 | .82 | .94 | .98 | 1.00 |

*Table D.3* Orthogonal polynomials for six occasions

| T1 | T2 | T3 | T4 |
|---|---|---|---|
| –5 | 5 | –5 | 1 |
| –3 | –1 | 7 | –3 |
| –1 | –4 | 4 | 2 |
| 1 | –4 | –4 | 2 |
| 3 | –1 | –7 | –3 |
| 5 | 5 | 5 | 1 |

square root of the sum of the squared values, which transforms all polynomials to the same scale, so they can be directly compared. The same effect can be accomplished by standardizing all orthogonal polynomials, which standardizes them to a mean of zero and a standard deviation of one.

Finally, if there are many polynomials, and the interest is only in controlling their combined effect, it is possible to combine them all into one predictor variable by using the predicted values in a polynomial regression model as the single predictor. This is known as the sheaf coefficient (Whitt, 1986).

# Author Index

# Subject Index