

2nd Edition

MULTILEVEL ANALYSIS

An
Introduction to
Basic and Advanced
Multilevel Modeling

Tom A B **SNIJDERS**
Roel J **BOSKER**





The second edition of this classic text introduces the main methods, techniques and issues involved in carrying out multilevel modeling and analysis.

SNIJDERS and **BOSKER'S** book is an applied, authoritative and accessible introduction to the topic, providing readers with a clear conceptual and practical understanding of all the main issues involved in designing multilevel studies and conducting multilevel analysis.

This book provides step-by-step coverage of:

- multilevel theories
- ecological fallacies
- the hierarchical linear model
- testing and model specification
- heteroscedasticity
- study designs
- longitudinal data
- multivariate multilevel models
- discrete dependent variables

There are also new chapters on:

- missing data
- multilevel modeling and survey weights
- Bayesian and MCMC estimation and latent-class models.

This book has been comprehensively revised and updated since the last edition, and now discusses modeling using HLM, MLwiN, SAS, Stata including GLLAMM, R, SPSS, Mplus, WinBugs, Latent Gold, and SuperMix.

This is a must-have text for any student, teacher or researcher with an interest in conducting or understanding multilevel analysis.

Tom A B **SNIJDERS**

is Professor of Statistics in the Social Sciences at the University of Oxford and Professor of Statistics and Methodology at the University of Groningen.

Roel J **BOSKER**

is Professor of Education and Director of GION, Groningen Institute for Educational Research, at the University of Groningen.

Cover image Stairs, La Paz, Bolivia © Sven Werkmeister

Cover design by Naomi C Robinson

ISBN-13: 978-1-84920-201-5



9 781849 202015



2nd Edition



MULTILEVEL ANALYSIS

An
Introduction to
Basic and Advanced
Multilevel Modeling

Tom A B **SNIJDERS**
Roel J **BOSKER**



Los Angeles | London | New Delhi
Singapore | Washington DC

© Tom A B Snijders and Roel J Bosker 2012

First edition published 1999

Reprinted 2002, 2003, 2004

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act, 1988, this publication may be reproduced, stored or transmitted in any form, or by any means, only with the prior permission in writing of the publishers, or in the case of reprographic reproduction, in accordance with the terms of licences issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publishers.

SAGE Publications Ltd

1 Oliver's Yard

55 City Road

London EC1Y 1SP

SAGE Publications Inc.

2455 Teller Road

Thousand Oaks, California 91320

SAGE Publications India Pvt Ltd

B 1/I 1 Mohan Cooperative Industrial Area

Mathura Road

New Delhi 110 044

SAGE Publications Asia-Pacific Pte Ltd

33 Pekin Street #02-01

Far East Square

Singapore 048763

Library of Congress Control Number: 2011926498

British Library Cataloguing in Publication data

A catalogue record for this book is available from the British Library

ISBN 978-1-84920-200-8

ISBN 978-1-84920-201-5 (pbk)

Typeset by C&M Digitals (P) Ltd, Chennai, India

Printed by MPG Books Group, Bodmin, Cornwall

Printed on paper from sustainable resources



Contents

Preface to the Second Edition	x
Preface to the First Edition	xii
1 Introduction	1
1.1 Multilevel analysis	1
1.1.1 Probability models	2
1.2 This book	3
1.2.1 Prerequisites	5
1.2.2 Notation	5
2 Multilevel Theories, Multistage Sampling, and Multilevel Models	6
2.1 Dependence as a nuisance	6
2.2 Dependence as an interesting phenomenon	8
2.3 Macro-level, micro-level, and cross-level relations	9
2.4 Glommary	13
3 Statistical Treatment of Clustered Data	14
3.1 Aggregation	14
3.2 Disaggregation	16
3.3 The intraclass correlation	17
3.3.1 Within-group and between-group variance	19
3.3.2 Testing for group differences	22
3.4 Design effects in two-stage samples	23
3.5 Reliability of aggregated variables	25
3.6 Within- and between-group relations	26
3.6.1 Regressions	27
3.6.2 Correlations	32
3.6.3 Estimation of within- and between-group correlations	34
3.7 Combination of within-group evidence	36
3.8 Glommary	39
4 The Random Intercept Model	41
4.1 Terminology and notation	42
4.2 A regression model: fixed effects only	43

4.3	Variable intercepts: fixed or random parameters?	44
4.3.1	When to use random coefficient models	46
4.4	Definition of the random intercept model	49
4.5	More explanatory variables	54
4.6	Within- and between-group regressions	56
4.7	Parameter estimation	60
4.8	'Estimating' random group effects: posterior means	62
4.8.1	Posterior confidence intervals	64
4.9	Three-level random intercept models	67
4.10	Glommary	71
5	The Hierarchical Linear Model	74
5.1	Random slopes	74
5.1.1	Heteroscedasticity	75
5.1.2	Do not force τ_{01} to be 0!	76
5.1.3	Interpretation of random slope variances	77
5.2	Explanation of random intercepts and slopes	80
5.2.1	Cross-level interaction effects	81
5.2.2	A general formulation of fixed and random parts	86
5.3	Specification of random slope models	87
5.3.1	Centering variables with random slopes?	87
5.4	Estimation	89
5.5	Three or more levels	90
5.6	Glommary	92
6	Testing and Model Specification	94
6.1	Tests for fixed parameters	94
6.1.1	Multiparameter tests for fixed effects	96
6.2	Deviance tests	97
6.2.1	More powerful tests for variance parameters	98
6.3	Other tests for parameters in the random part	100
6.3.1	Confidence intervals for parameters in the random part	100
6.4	Model specification	102
6.4.1	Working upward from level one	104
6.4.2	Joint consideration of level-one and level-two variables	106
6.4.3	Concluding remarks on model specification	107
6.5	Glommary	108
7	How Much Does the Model Explain?	109
7.1	Explained variance	109
7.1.1	Negative values of R^2 ?	110
7.1.2	Definitions of the proportion of explained variance in two-level models	111
7.1.3	Explained variance in three-level models	113
7.1.4	Explained variance in models with random slopes	113

7.2	Components of variance	114
7.2.1	Random intercept models	114
7.2.2	Random slope models	116
7.3	Glommary	117
8	Heteroscedasticity	119
8.1	Heteroscedasticity at level one	119
8.1.1	Linear variance functions	119
8.1.2	Quadratic variance functions	123
8.2	Heteroscedasticity at level two	128
8.3	Glommary	129
9	Missing Data	130
9.1	General issues for missing data	131
9.1.1	Implications for design	133
9.2	Missing values of the dependent variable	133
9.3	Full maximum likelihood	134
9.4	Imputation	135
9.4.1	The imputation method	137
9.4.2	Putting together the multiple results	140
9.5	Multiple imputations by chained equations	144
9.6	Choice of the imputation model	148
9.7	Glommary	149
10	Assumptions of the Hierarchical Linear Model	152
10.1	Assumptions of the hierarchical linear model	153
10.2	Following the logic of the hierarchical linear model	154
10.2.1	Include contextual effects	154
10.2.2	Check whether variables have random effects	155
10.2.3	Explained variance	156
10.3	Specification of the fixed part	156
10.4	Specification of the random part	158
10.4.1	Testing for heteroscedasticity	159
10.4.2	What to do in case of heteroscedasticity	161
10.5	Inspection of level-one residuals	161
10.6	Residuals at level two	165
10.7	Influence of level-two units	167
10.8	More general distributional assumptions	172
10.9	Glommary	173
11	Designing Multilevel Studies	176
11.1	Some introductory notes on power	177
11.2	Estimating a population mean	179
11.3	Measurement of subjects	180
11.4	Estimating association between variables	180
11.4.1	Cross-level interaction effects	185
11.5	Allocating treatment to groups or individuals	187

11.6 Exploring the variance structure	188
11.6.1 The intraclass correlation	188
11.6.2 Variance parameters	190
11.7 Glommary	191
12 Other Methods and Models	194
12.1 Bayesian inference	194
12.2 Sandwich estimators for standard errors	197
12.3 Latent class models	201
12.4 Glommary	203
13 Imperfect Hierarchies	205
13.1 A two-level model with a crossed random factor	206
13.2 Crossed random effects in three-level models	210
13.3 Multiple membership models	210
13.4 Multiple membership multiple classification models	213
13.5 Glommary	215
14 Survey Weights	216
14.1 Model-based and design-based inference	217
14.1.1 Descriptive and analytic use of surveys	217
14.2 Two kinds of weights	219
14.3 Choosing between model-based and design-based analysis	222
14.3.1 Inclusion probabilities and two-level weights	223
14.3.2 Exploring the informativeness of the sampling design	225
14.4 Example: Metacognitive strategies as measured in the PISA study	231
14.4.1 Sampling design	231
14.4.2 Model-based analysis of data divided into parts	233
14.4.3 Inclusion of weights in the model	235
14.5 How to assign weights in multilevel models	236
14.6 Appendix. Matrix expressions for the single-level estimators	244
14.7 Glommary	244
15 Longitudinal Data	247
15.1 Fixed occasions	248
15.1.1 The compound symmetry model	249
15.1.2 Random slopes	253
15.1.3 The fully multivariate model	255
15.1.4 Multivariate regression analysis	260
15.1.5 Explained variance	261
15.2 Variable occasion designs	263
15.2.1 Populations of curves	263
15.2.2 Random functions	263
15.2.3 Explaining the functions	274
15.2.4 Changing covariates	276
15.3 Autocorrelated residuals	280
15.4 Glommary	280

16 Multivariate Multilevel Models	282
16.1 Why analyze multiple dependent variables simultaneously?	283
16.2 The multivariate random intercept model	283
16.3 Multivariate random slope models	288
16.4 Glommary	288
17 Discrete Dependent Variables	289
17.1 Hierarchical generalized linear models	289
17.2 Introduction to multilevel logistic regression	290
17.2.1 Heterogeneous proportions	290
17.2.2 The logit function: Log-odds	293
17.2.3 The empty model	295
17.2.4 The random intercept model	297
17.2.5 Estimation	300
17.2.6 Aggregation	301
17.3 Further topics on multilevel logistic regression	302
17.3.1 Random slope model	302
17.3.2 Representation as a threshold model	303
17.3.3 Residual intraclass correlation coefficient	304
17.3.4 Explained variance	305
17.3.5 Consequences of adding effects to the model	307
17.4 Ordered categorical variables	310
17.5 Multilevel event history analysis	313
17.6 Multilevel Poisson regression	314
17.7 Glommary	320
18 Software	323
18.1 Special software for multilevel modeling	323
18.1.1 HLM	324
18.1.2 MLwiN	325
18.1.3 The MIXOR suite and SuperMix	325
18.2 Modules in general-purpose software packages	327
18.2.1 SAS procedures VARCOMP, MIXED, GLIMMIX, and NLMIXED	327
18.2.2 R	328
18.2.3 Stata	328
18.2.4 SPSS, commands VARCOMP and MIXED	329
18.3 Other multilevel software	329
18.3.1 PinT	329
18.3.2 Optimal Design	330
18.3.3 MLPowSim	330
18.3.4 Mplus	330
18.3.5 Latent Gold	330
18.3.6 REALCOM	330
18.3.7 WinBUGS	331
References	332
Index	348

Preface to the Second Edition

Multilevel analysis is a domain of data analysis that has been developing strongly before as well as after the publication of the first edition of our book in 1999. This second edition has been seriously revised. It contains more material, it has been updated and corrected, and a number of explanations were clarified.

The main new material is in three new chapters. A chapter was added about missing data, and another about the use of multilevel modeling for surveys with nonconstant inclusion probabilities ('survey weights'). Also a chapter was added in which three special techniques are briefly treated: Bayesian estimation, cluster-robust standard errors (the so-called sandwich standard error), and latent class models. The topics of these new chapters all belong to the 'advanced' part of the title. Among what was not covered in the first edition, these are some of the topics which we believe are most frequently encountered in the practice of multilevel research.

New material has been added also to existing chapters. The main example (starting in Chapter 4) has been renewed because the treatment of missing data in the old version was inadequate. Various other new examples also have been added. Further, there now is a more elaborate treatment of the combination of within-group evidence without using full-blown multilevel modeling (Section 3.7); more detailed considerations are discussed for choosing between fixed and random effects models (Section 4.3); diagnostic and comparative standard errors of posterior means are explained (Section 4.8.1); the treatment of tests for parameters of the random part was corrected and confidence intervals for these parameters are discussed (Sections 6.2 and 6.3); multiple membership models are treated in Chapter 13; and there has been an overhaul of the treatment of estimation methods for hierarchical generalized linear models in Chapter 17. Chapter 18 about multilevel software was totally rewritten, keeping it relatively short because this is the part of any textbook that ages most rapidly. Throughout all chapters new developments have been mentioned, pointers are given to the recent literature, various difficulties now are explained in more elaborate ways, and errors have been corrected.

All chapters (from the second on) now start by an overview, and are concluded (except for the last) by a 'glommary'. As every reader will know after reading this book, this is a summary of the main concepts treated in the chapter in a form akin to a glossary. Our intention is that these new elements will improve the didactical qualities of this textbook. Having said this, we think that the understanding of the book, or parts of it, may be further enhanced by going through the examples using the data that we made available (as far as this was allowed) at <http://www.stats.ox.ac.uk/~snijders/mlbook.htm>. This website will also contain our comments on remarks made on the book by industrious readers, as well as our corrections for errors if any will be discovered.

We are very grateful for stimulating discussions (over the years or in the recent period of revising the text), comments on drafts of chapters, and help with software, to many people: Marnie Bertolet, sir David Cox, Roel de Jong, Jon Fahlander, Mark Huisman, Johan Koskinen, Catalina Lomos, Mayra Mascareño, Paulina Preciado, Roy Stewart, Anneke Timmermans, and Marijtje van Duijn. For help with new data sets we are grateful to some of these people and also to Hennie Brandsma, Simone Doolaard, Sonja Drobnič, Anja Knuver, Hans Kuyper, Sascha Peter, Stijn Ruiter, Greetje van der Werf, and Frank van Tubergen.

Tom Snijders

Roel Bosker

March, 2011

Preface to the First Edition

This book grew out of our teaching and consultation activities in the domain of multilevel analysis. It is intended for the absolute beginner in this field as well as for those who have already mastered the fundamentals and are now entering more complicated areas of application. The reader is referred to Section 1.2 for an overview of this book and for some reading guidelines.

We are grateful to various people from whom we got reactions on earlier parts of this manuscript and also to the students who were exposed to it and helped us realize what was unclear. We received useful comments from, and benefited from discussions about parts of the manuscript with, among others, Joerg Blasius, Marijtje van Duijn, Wolfgang Langer, Ralf Maslowski, and Ian Plewis. Moreover we would like to thank Hennie Brandsma, Mieke Brekelmans, Jan van Damme, Hetty Dekkers, Miranda Lubbers, Lyset Rekers-Mombarg and Jan Maarten Wit, Carolina de Weerth, Beate Völker, Ger van der Werf, and the Zentral Archiv (Cologne) who kindly permitted us to use data from their respective research projects as illustrative material for this book. We would also like to thank Annelies Verstappen-Remmers for her unfailing secretarial assistance.

*Tom Snijders
Roel Bosker
June, 1999*

1

Introduction

1.1 Multilevel analysis

Multilevel analysis is a methodology for the analysis of data with complex patterns of variability, with a focus on nested sources of such variability – pupils in classes, employees in firms, suspects tried by judges in courts, animals in litters, longitudinal measurements of subjects, etc. In the analysis of such data, it is usually illuminating to take account of the fact that each level of nesting is associated with variability that has a distinct interpretation. There is variability, for example, between pupils but also between classes, and one may draw incorrect conclusions if no distinction is made between these different sources of variability. Multilevel analysis is an approach to the analysis of such data, including the statistical techniques as well as the methodology for their use. The term ‘multilevel analysis’ is used mainly in the social sciences (in the wide sense: sociology, education, psychology, economics, criminology, etc.), but also in other fields such as the biomedical sciences. Our focus will be on the social sciences. Other terms, referring to the technical aspects, are hierarchical linear models, mixed models, and random coefficient models.

In its present form, multilevel analysis is a stream which has two tributaries: contextual analysis and mixed effects models. *Contextual analysis* is a development in the social sciences which has focused on the effects of the social context on individual behavior. Some landmarks before 1980 are the paper by Robinson (1950) which discussed the ecological fallacy (which refers to confusion between aggregate and individual effects), the paper by Davis et al. (1961) on the distinction between within-group and between-group regression, the volume edited by Dogan and Rokkan (1969), and the paper by Burstein et al. (1978) on treating regression intercepts and slopes on one level as outcomes on the higher level.

Mixed effects models are statistical models in the analysis of variance (ANOVA) and in regression analysis where it is assumed that some of the coefficients are fixed and others are random. This subject is too vast even to mention some landmarks. A standard reference book on random effects models and mixed effects models is Searle et al. (1992), Chapter 2 of which gives an extensive historical overview. The name ‘mixed model’ seems to have been used first by Eisenhart (1947).

Contextual modeling until about 1980 focused on the definition of appropriate variables to be used in ordinary least squares regression analysis. Until the 1980s the main

focus in the development of statistical procedures for mixed models was on random effects (i.e., random differences between categories in some classification system) more than on random coefficients (i.e., random effects of numerical variables). Multilevel analysis as we now know it was formed by these two streams coming together. It was realized that, in contextual modeling, the individual and the context are distinct sources of variability, which should both be modeled as random influences. On the other hand, statistical methods and algorithms were developed that allowed the practical use of regression-type models with nested random coefficients. There was a cascade of statistical papers: Aitkin et al. (1981), Laird and Ware (1982), Mason et al. (1983), Goldstein (1986), Aitkin and Longford (1986), Raudenbush and Bryk (1986), de Leeuw and Kreft (1986), and Longford (1987) proposed and developed techniques for calculating estimates for mixed models with nested coefficients. These techniques, together with the programs implementing them which were developed by a number of these researchers or under their supervision, allowed the practical use of models of which until that moment only special cases were accessible for practical use. By 1986 the basis of multilevel analysis was established. The first textbook appeared (by Goldstein, now in its fourth edition) and was followed by a few others in the 1990s and many more in the 2000s. The methodology has been further elaborated since then, and has proved to be quite fruitful in applications in many fields. On the organizational side, there are stimulating centers such as the ‘Multilevel Models Project’ in Bristol with its Newsletter and its website <http://www.mlwin.com/>, and there is an active internet discussion group at <http://www.jiscmail.ac.uk/lists/multilevel.html>.

In the biomedical sciences mixed models were proposed especially for longitudinal data; in economics mainly for panel data (Swamy, 1971), the most common longitudinal data in economics. One of the issues treated in the economics literature was the pooling of cross-sectional and time series data (e.g., Maddala, 1971; Hausman and Taylor, 1981), which is closely related to the difference between within-group and between-group regressions. Overviews are given by Chow (1984) and Baltagi (2008).

A more elaborate history of multilevel analysis is presented in the bibliographical sections of Longford (1993) and in Kreft and de Leeuw (1998). For an extensive bibliography of the older literature, see Hüttner and van den Eeden (1995). A more recent overview of much statistical work in this area can be found in the handbook by de Leeuw and Meijer (2008a).

1.1.1 Probability models

The main statistical model of multilevel analysis is the hierarchical linear model, an extension of multiple linear regression to a model that includes nested random coefficients. This model is explained in Chapter 5 and forms the basis of most of this book.

There are several ways to argue why it makes sense to use a probability model for data analysis. In sampling theory a distinction is made between *design-based inference* and *model-based inference*. This is discussed further in Chapter 14. The former means that the researcher draws a probability sample from some finite population, and wishes to make inferences from the sample to this finite population. The probability model then follows from how the sample is drawn by the researcher. Model-based inference means that the researcher postulates a probability model, usually aiming at inference to some large and sometimes hypothetical population such as all English primary school pupils in the 2000s or all human adults living in a present-day industrialized culture. If the probability model

is adequate then so are the inferences based on it, but checking this adequacy is possible only to a limited extent.

It is possible to apply model-based inference to data collected by investigating some entire research population, such as all 12-year-old pupils in Amsterdam at a given moment. Sometimes the question arises as to why one should use a probability model if no sample is drawn but an entire population is observed. Using a probability model that assumes statistical variability, even though an entire research population was investigated, can be justified by realizing that conclusions are sought which apply not only to the investigated research population but also to a wider population. The investigated research population is assumed to be representative of this wider population – for pupils who are older or younger, in other towns, perhaps in other countries. This is called a *superpopulation* in Chapter 14, where the relation between model-based and design-based inference is further discussed. Applicability of the statistical inference to such a wider population is not automatic, but has to be carefully argued by considering whether indeed the research population may be considered to be representative of the larger (often vaguely outlined) population. This is the ‘second span of the bridge of statistical inference’ as discussed by Cornfield and Tukey (1956).¹ The inference then is not primarily about a given delimited set of individuals but about social, behavioral, biological, etc., mechanisms and processes. The random effects, or residuals, playing a role in such probability models can be regarded as resulting from the factors that are not included in the explanatory variables used. They reflect the approximative nature of the model used. The model-based inference will be adequate to the extent that the assumptions of the probability model are an adequate reflection of the effects that are not explicitly included by means of observed variables.

As we shall see in Chapters 3–5, the basic idea of multilevel analysis is that data sets with a nesting structure that includes unexplained variability at each level of nesting, such as pupils in classes or employees in firms, are usually not adequately represented by the probability model of multiple linear regression analysis, but are often adequately represented by the hierarchical linear model. Thus, the use of the hierarchical linear model in multilevel analysis is in the tradition of model-based inference.

1.2 This book

This book is intended as an introductory textbook and as a reference book for practical users of multilevel analysis. We have tried to include all the main points that come up when applying multilevel analysis. Most of the data sets used in the examples, and corresponding commands to run the examples in various computer programs (see Chapter 18), are available on the website <http://www.stats.ox.ac.uk/~snijders/mlbook.htm>.



After this introductory chapter, the book proceeds with a conceptual chapter about multilevel questions and a chapter on ways to treat multilevel data that are not based on the hierarchical linear model. Chapters 4–6 treat the basic conceptual ideas of the hierarchical linear model, and how to work with it in practice. Chapter 4 introduces the random intercept model as the primary example of the hierarchical linear model. This is extended in Chapter 5 to random slope models. Chapters 4 and 5 focus on understanding the hierarchical linear model and its parameters, paying only very limited attention to procedures

¹We are indebted to Ivo Molenaar for this reference.

and algorithms for parameter estimation (estimation being work that most researchers delegate to the computer). Chapter 6 is concerned with testing parameters and specifying a multilevel model.

An introductory course on multilevel analysis could cover Chapters 1–6 and Section 7.1, with selected material from other chapters. A minimal course would focus on Chapters 4–6. The later chapters cover topics that are more specialized or more advanced, but important in the practice of multilevel analysis.

The text of this book is not based on a particular computer program for multilevel analysis. The last chapter, 18, gives a brief review of computer programs that can be used for multilevel analysis.

Chapters 7 (on the explanatory power of the model) and 10 (on model assumptions) are important for the interpretation of the results of statistical analyses using the hierarchical linear model. Researchers who have data sets with many missing values, or who plan to collect data sets that may run this type of risk, will profit from reading Chapter 9. Chapter 11 helps the researcher in setting up a multilevel study, and in choosing sample sizes at the various levels.

Some multilevel data sets come from surveys done according to a complex design, associated with survey weights reflecting the undersampling and oversampling of parts of the population. Ways to analyze such data sets using the hierarchical linear model are covered in Chapter 14.

Several methods and models have been developed that can sometimes be useful as additions or alternatives to the more commonly used methods for the hierarchical linear model. Chapter 12 briefly presents three of these: Bayesian procedures, the sandwich estimator for standard errors, and latent class models.

Chapters 8 and 13–17 treat various extensions of the basic hierarchical linear model that are useful in practical research. The topic of Chapter 8, heteroscedasticity (nonconstant residual variances), may seem rather specialized. Modeling heteroscedasticity, however, is easily done within the hierarchical linear model and can be very useful. It also allows model checks and model modifications that are used in Chapter 10. Chapter 13 treats data structures where the different sources of variability, the ‘levels’ of the multilevel analysis, are not nested but related in different ways: crossed classifications and multiple memberships. Chapter 15 is about longitudinal data, with a fixed occasion design (i.e., repeated measures data) as well as those with a variable occasion design, where the time moments of measurement may differ arbitrarily between subjects. This chapter indicates how the flexibility of the multilevel model gives important opportunities for data analysis (e.g., for incomplete multivariate or longitudinal data) that were unavailable earlier. Chapter 16 is about multilevel analysis for multivariate dependent variables. Chapter 17 describes the multilevel modeling of dichotomous, ordinal, and frequency data.

Each chapter starts with an overview and finishes with a summarizing glossary, which we have called a *glommary*. The glommaries can be consulted to gain rapid overviews of what is treated in the various chapters.

If additional textbooks are sought, one could consider the excellent introductions by Hox (2010) and Gelman and Hill (2007); Raudenbush and Bryk (2002), for an elaborate treatment of the hierarchical linear model; Longford (1993), Goldstein (2011), Demidenko (2004), and de Leeuw and Meijer (2008a) for more detailed mathematical background; and Skrondal and Rabe-Hesketh (2004) for further modeling, especially latent variable models.

1.2.1 Prerequisites

In order to read this textbook, a good working knowledge of statistics is required. It is assumed that you know the concepts of probability, random variable, probability distribution, population, sample, statistical independence, expectation (population mean), variance, covariance, correlation, standard deviation, and standard error. Furthermore, it is assumed that you know the basics of hypothesis testing and multiple regression analysis, and that you can understand formulas of the kind that occur in the explanation of regression analysis.

Matrix notation is used only in a few more advanced sections. These sections can be skipped without loss of understanding of other parts of the book.

1.2.2 Notation

The main notational conventions are as follows. Abstract variables and random variables are denoted by italicized capital letters, such as X or Y . Outcomes of random variables and other fixed values are denoted by italicized lower-case letters, such as x or z . Thus we speak about the variable X , but in formulas where the value of this variable is considered as a fixed, nonrandom value, it will be denoted by x . There are some exceptions to this, for example in Chapter 2 and in the use of the letter N for the number of groups ('level-two units') in the data.

The letter \mathcal{E} is used to denote the *expected value*, or population average, of a random variable. Thus, $\mathcal{E}Y$ and $\mathcal{E}(Y)$ denote the expected value of Y . For example, if P_n is the fraction of tails obtained in n coin flips, and the coin is fair, then the expected value is $\mathcal{E}P_n = \frac{1}{2}$.

Statistical parameters are indicated by Greek letters. Examples are μ , σ^2 , and β . The following Greek letters are used.

α	alpha
β	beta
γ	gamma
δ	delta
η	eta
θ	theta
λ	lambda
μ	mu
π	pi
ρ	rho
σ	sigma
τ	tau
φ	phi
χ	chi
ω	omega
Δ	capital Delta
Σ	capital Sigma
T	capital Tau
X	capital Chi

2

Multilevel Theories, Multistage Sampling, and Multilevel Models

Phenomena and data sets in the social sciences often have a multilevel structure. This may be reflected in the design of data collection: simple random sampling is often not a very cost-efficient strategy, and multistage samples may be more efficient instead. This chapter is concerned with the reasons why it is important to take account of the clustering of the data, also called their multilevel structure, in the data analysis phase.

OVERVIEW OF THE CHAPTER

First we discuss how methods of inference failing to take into account the multilevel data structure may lead to erroneous conclusions, because independence assumptions are likely to be violated. The next two sections sketch the reasons for interest in a multilevel approach from the applications point of view. In many cases the multilevel data structure reflects essential aspects of the social (biological, etc.) world, and important research questions can be formulated about relations between variables at different layers in a hierarchical system. In this case the dependency of observations within clusters is of focal interest, because it reflects the fact that clusters differ in certain respects. In either case, the use of single-level statistical models is no longer valid. The fallacies to which their use can lead are described in the next chapter.

2.1 Dependence as a nuisance

Textbooks on statistics tell us that observations should be sampled *independently* of each other as standard. Thus the standard sampling design on which statistical models are based is simple random sampling with replacement from an infinite population: the result of one selection is independent of the result of any other selection, and all single units in the population have the same chances of being selected into the sample.

Textbooks on sampling, however, make it clear that there are more cost-efficient sampling designs, based on the idea that probabilities of selection should be known but do not have to be constant. One of those cost-efficient sampling designs is the *multistage sample*: the population of interest consists of subpopulations, also called *clusters*, and selection takes place via those subpopulations.

If there is only one subpopulation level, the design is a *two-stage sample*. Pupils, for instance, are grouped in schools, so the population of pupils consists of subpopulations of schools that contain pupils. Other examples are: families in neighborhoods, teeth in jawbones, animals in litters, employees in firms, and children in families. In a random two-stage sample, a random sample of the primary units (schools, neighborhoods, jawbones, litters, firms, families) is taken in the first stage, and then the secondary units (pupils, families, teeth, animals, employees, children) are sampled at random from the selected primary units in the second stage. A common mistake in research is to ignore the fact that the sampling scheme was a two-stage one, and to pretend that the secondary units were selected independently. The mistake in this case would be that the researcher is overlooking the fact that the secondary units were not sampled independently of each other: having selected a primary unit (e.g., a school) increases the chances of selection of secondary units (e.g., pupils) from that primary unit. In other words, the multistage sampling design leads to *dependent* observations, and failing to deal with this properly in the statistical analysis may lead to erroneous inferences. An example of the grossly inflated type I error rates that may then occur is given by Dorman (2008).

The multistage sampling design can be depicted graphically as in Figure 2.1. This shows a population that consists of 10 subpopulations, each containing 10 micro-units. A sample of 25% is taken by randomly selecting 5 out of 10 subpopulations and within these – again at random of course – 5 out of 10 micro-units.

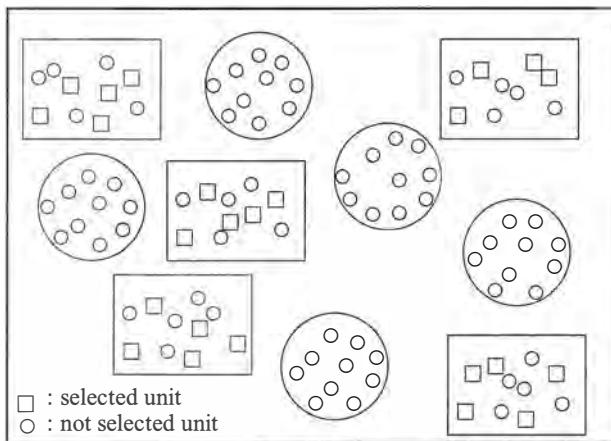


Figure 2.1: Multistage sample.

Multistage samples are preferred in practice, because the costs of interviewing or testing persons are reduced enormously if these persons are geographically or organizationally grouped. Such sample designs correspond to the organization of the social world. It is cheaper to travel to 100 neighbourhoods and interview 10 persons per neighbourhood on

their political preferences than to travel to 1,000 neighbourhoods and interview one person per neighbourhood. In the next chapters we will see how we can make adjustments to deal with these dependencies.

2.2 Dependence as an interesting phenomenon

The previous section implies that, if we want to make inferences on, for example, the earnings of employees in the for-profit sector, it is cost-efficient to use a multistage sampling design in which employees are selected via the firms in which they work. A common feature in social research, however, is that in many cases we wish to make inferences on the firms as well as on the employees. Questions that we seek to answer may be: Do employees in multinationals earn more than employees in other firms? Is there a relation between the performance of pupils and the experience of their teacher? Is the sentence differential between black and white suspects different between judges, and if so, can we find characteristics of judges to which this sentence differential is related? In this case a variable is defined at the primary unit level (firms, teachers, judges) as well as at the secondary unit level (employees, pupils, cases). Henceforth we will refer to primary units as *macro-level units* (or macro-units for short) and to secondary units as *micro-level units* (or micro-units for short). The micro level is called the *lower level* (first) and the macro level is called the *higher level* (second). For the time being, we will restrict ourselves to the two-level case, and thus to two-stage samples only. Table 2.1 gives a summary of the terminology.

Table 2.1: Summary of terms to describe units at either level in the two-level case.

macro-level units	micro-level units
macro-units	micro-units
primary units	secondary units
clusters	elementary units
level-two units	level-one units

Examples of macro-units and the micro-units nested within them are presented in Table 2.2. Most of the examples presented in the table have been dealt with in the text already. It is important to note that what is defined as a macro-unit or a micro-unit depends on the theory at hand. Teachers are nested within schools, if we study organizational effects on teacher burn-out then teachers are the micro-units and schools the macro-units. But when studying teacher effects on student achievement, teachers are the macro-units and students the micro-units. The same goes, *mutatis mutandis*, for neighborhoods and families (e.g., when studying the effects of housing conditions on marital problems), and for families and children (e.g., when studying effects of income on educational performance of siblings).

In all these instances the dependency of the observations on the micro-units within the macro-units is of focal interest. If we stick to the example of schools and pupils, then the dependency (e.g., in mathematics achievement of pupils within a school) may stem from:

Table 2.2: Some examples of units at the macro and micro level.

Macro level	Micro level
schools	teachers
classes	pupils
neighbourhoods	families
firms	employees
jawbones	teeth
families	children
litters	animals
doctors	patients
subjects	measurements
interviewers	respondents
judges	suspects

1. pupils within a school sharing the same school environment;
2. pupils within a school sharing the same teachers;
3. pupils within a school affecting each other by direct communication or shared group norms;
4. pupils within a school coming from the same neighborhood.

The more the achievement levels of pupils within a school are alike (as compared to pupils from other schools), the more likely it is that causes of the achievement have to do with the organizational unit (in this case, the school). Absence of dependency in this case implies absence of institutional effects on individual performance.

A special kind of nesting is defined by longitudinal data, represented in Table 2.2 as ‘measurements within subjects’. The measurement occasions here are the micro-units and the subjects the macro-units. The dependence of the different measurements for a given subject is of primary importance in longitudinal data, but the following section on relations between variables defined at either level is not directly intended for the nesting structure defined by longitudinal data. Because of the special nature of this nesting structure, Chapter 15 is specifically devoted to it.

The models treated in this book are for situations where the dependent variable is at the lowest level. For models with nested data sets where the dependent variable is defined at a higher level one may consult Croon and van Veldhoven (2007), Lüdtke et al. (2008), and van Mierlo et al. (2009).

2.3 Macro-level, micro-level, and cross-level relations

In the study of hierarchical or multilevel systems, Lazarsfeld and Menzel (1971) made important distinctions between properties and propositions connected to the different levels.

In his summary of this work, Tacq (1986) distinguished between three kinds of propositions: on micro-units (e.g., ‘employees have on average 4 effective working hours per day’; ‘boys lag behind girls in reading comprehension’), on macro-units (e.g., ‘schools have on average a budget of \$20,000 to spend on resources’; ‘in neighborhoods with bad housing conditions crime rates are above average’), or on macro-micro relations (e.g., ‘if firms have a salary bonus system, the productivity of employees will be greater’; ‘a child suffering from a broken family situation will affect the climate in the classroom’).

Multilevel statistical models are always¹ called for if we are interested in propositions that connect variables defined at different levels, the micro and the macro, and also if a multistage sample design has been employed. The use of such a sampling design is quite obvious if we are interested in macro-micro relations, less obvious (but often necessary from a cost-effectiveness point of view) if micro-level propositions are our primary concern, and hardly obvious at all (but sometimes still applicable) if macro-level propositions are what we are focusing on. These three instances will be discussed below. To facilitate comprehension, following Tacq (1986) we use figures with the following conventions: a dotted line indicates that there are two levels; below the line is the micro level; above the line is the macro level; macro-level variables are denoted by capitals; micro-level variables are denoted by lower-case letters; and arrows denote presumed causal relations.

Multilevel propositions

Multilevel propositions can be represented as in Figure 2.2. In this example we are interested in the effect of the macro-level variable Z (e.g., teacher efficacy) on the micro-level variable y (e.g., pupil motivation), controlling for the micro-level variable x (e.g., pupil aptitude).

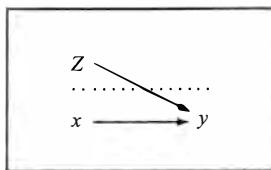


Figure 2.2: The structure of a multilevel proposition.

Micro-level propositions

Micro-level propositions are of the form indicated in Figure 2.3. In this case the line indicates that there is a macro level which is not referred to in the hypothesis that is put to the test, but which is used in the sampling design in the first stage. In assessing the strength of the relation between occupational status and income, for instance, respondents may have been selected for face-to-face interviews by zip-code area. This then may cause dependency (as a nuisance) in the data.

¹As with any rule, there are exceptions. If the data set is such that for each macro-unit only one micro-unit is included in the sample, single-level methods still can be used.

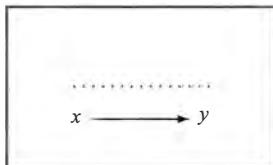


Figure 2.3: The structure of a micro-level proposition.

Macro-level propositions

Macro-level propositions are of the form of Figure 2.4. The line separating the macro level from the micro level seems superfluous here. When investigating the relation between the long-range strategic planning policy of firms and their profits, there is no multilevel situation, and a simple random sample may have been taken. When either or both variables are not directly observable, however, and have to be measured at the micro level (e.g., organizational climate measured as the average satisfaction of employees), then a two-stage sample is needed nevertheless. This is the case *a fortiori* for variables defined as aggregates of micro-level variables (e.g., the crime rate in a neighborhood).

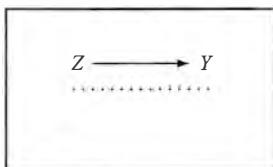


Figure 2.4: The structure of a macro-level proposition.

Macro-micro relations

The most common situation in social research is that macro-level variables are supposed to have a relation with micro-level variables. There are three obvious instances of macro-to-micro relations, all of which are typical examples of the multilevel situation (see Figure 2.5). The first case is the macro-to-micro proposition. The more explicit the religious norms in social networks, for example, the more conservative the views that individuals have on contraception. The second proposition is a special case of this. It refers to the case where there is a relation between Z and y , given that the effect of x on y is taken into account. The example given may be modified to: ‘for individuals of a given educational level’. The last case in the figure is the *macro-micro-interaction*, also known as the cross-level interaction: the relation between x and y is dependent on Z . To put it another way, the relation between Z and y is dependent on x . The effect of aptitude on achievement, for instance, may be small in case of ability grouping of pupils within classrooms but large in ungrouped classrooms.

Next to these three situations there is the so-called emergent, or micro-macro, proposition (Figure 2.6). In this case, a micro-level variable x affects a macro-level variable Z (student achievement may affect teachers’ experience of stress).

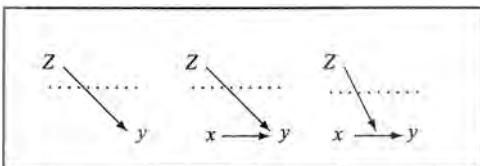


Figure 2.5: The structure of macro–micro propositions.

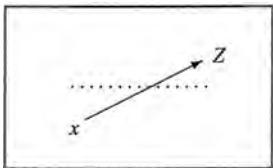


Figure 2.6: The structure of a micro–macro proposition.

It is of course possible to form combinations of the various examples given. Figure 2.7 contains a causal chain that explains through which micro-variables there is an association between the macro-level variables W and Z (cf. Coleman, 1990). As an example of this chain, we may be interested in why the qualities of a football coach affect his social prestige. The reason is that good coaches are capable of motivating their players, thus leading the players to good performance, thus to winning games, and this of course leads to more social prestige for the coach. Another instance of a complex multilevel proposition is the contextual effects proposition. For example, low socio-economic status pupils achieve less in classrooms with a low average aptitude. This is also a cross-level interaction effect, but the macro-level variable, average aptitude in the classroom, is now an aggregate of a micro-level variable.

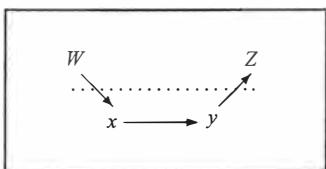


Figure 2.7: A causal macro–micro–micro–macro chain.

The methodological advances in multilevel modeling are now also leading to theoretical advances in contextual research: suitable definitions of context and ‘levels’, meaningful ways of aggregating variables to higher levels, conceptualizing and analyzing the interplay between characteristics of lower- and higher-level units. Some examples in various disciplines are the following. Following up on the initial work of Hauser (1970, 1974), in which he stated that group composition effects may be artifacts of underspecification of the micro-level model, Harker and Tymms (2004) discuss the issue of group composition effects in education. Sampson et al. (2002) give a review of theoretical work in the analysis of neighborhood effects. Diez-Roux (2000), Blakely and Woodward (2000), and O’Campo (2003) comment on advances along these lines in epidemiology and public health.

In the next chapters the statistical tools to handle multilevel structures will be introduced for outcome variables defined at the micro level.

2.4 Glommary

Multilevel data structures. Many data sets in the social sciences have a multilevel structure, that is, they constitute hierarchically nested systems with multiple levels. Much of our discussion focuses on two-level structures, but this can be generalized to three or more nested levels.

Sampling design. Often the multilevel nature of the social world leads to the practical efficiency of multistage samples. The population then consists of a nested system of subpopulations, and a nested sample is drawn accordingly. For example, when employing a random two-stage sample design, in the first stage a random sample of the primary units is taken, and in the second stage the secondary units are sampled at random from the selected primary units.

Levels. The levels are numbered such that the most detailed level is the first. For example, in a two-level structure of individuals nested in groups the individuals are called level-one units and the groups level-two units. (Note the different terminology compared to the words used in theories of survey sampling: in the preceding example, the primary units are the level-two units and the secondary units the level-one units.)

Units. The elements of a level are called units. Higher-level units are also called clusters. We talk about level-one units, level-two units, etc.

Dependence as a nuisance. Not taking account of the multilevel data structure, or of the multistage sampling design, is likely to lead to the use of statistical procedures in which independence assumptions are violated so that conclusions may be unfounded.

Dependence as an interesting phenomenon. The importance of the multilevel structure of social (biological, etc.) reality implies that research can often become more interesting when it takes account of the multilevel structure.

Multilevel propositions. Illustrations were given of scientific propositions involving multiple levels: micro-level propositions, macro-level propositions, macro–micro relations, cross-level interaction, and emergent propositions or micro–macro relations.

3

Statistical Treatment of Clustered Data

Before proceeding in the next chapters to the explanation of the hierarchical linear model, the main statistical model for multilevel analysis, this chapter looks at approaches to analyzing multilevel data sets that are more elementary and do not use this model.

OVERVIEW OF THE CHAPTER

The chapter starts by considering what will happen if we ignore the multilevel structure of the data. Are there any instances where one may proceed with single-level statistical models although the data stem from a multistage sampling design? What kind of errors – so-called ecological fallacies – may occur when this is done? The rest of the chapter is devoted to some statistical methods for multilevel data that attempt to uncover the role played by the various levels without fitting a full-blown hierarchical linear model. First, we describe the intraclass correlation coefficient, a basic measure for the degree of dependency in clustered observations. Second, some simple statistics (mean, standard error of the mean, variance, correlation, reliability of aggregates) are treated for two-stage sampling designs. To avoid ecological fallacies it is essential to distinguish within-group from between-group regressions. These concepts are explained, and the relations are spelled out between within-group, between-group, and total regressions, and similarly for correlations. Finally, we mention some simple methods for combining evidence from several independent studies, or groups, in a combined test or a combined estimate.

3.1 Aggregation

A common procedure in social research with two-level data is to aggregate the micro-level data to the macro level. The simplest way to do this is to work with the averages for each macro-unit.

There is nothing wrong with aggregation in cases where the researcher is only interested in macro-level propositions, although it should be borne in mind that the reliability of an aggregated variable depends, *inter alia*, on the number of micro-level units in a macro-level unit (see later in this chapter), and thus will be larger for the larger macro-units than for

the smaller ones. In cases where the researcher is interested in macro–micro or micro-level propositions, however, aggregation may result in gross errors.

The first potential error is the ‘shift of meaning’ (cf. Firebaugh, 1978; Hüttner, 1981). A variable that is aggregated to the macro level refers to the macro-units, not directly to the micro-units. The firm average of an employee rating of working conditions, for example, may be used as an index for ‘organizational climate’. This variable refers to the firm, not directly to the employees.

The second potential error with aggregation is the ecological fallacy (Robinson, 1950). A correlation between macro-level variables cannot be used to make assertions about micro-level relations. The percentage of black inhabitants in a neighborhood could be related to average political views in the neighborhood – for example, the higher the percentage of blacks in a neighborhood, the higher might be the proportion of people with extreme right-wing political views. This, however, does not give us any clue about the micro-level relation between race and political conviction. (The shift of meaning plays a role here, too. The percentage of black inhabitants is a variable that means something for the neighborhood, and this meaning is distinct from the meaning of ethnicity as an individual-level variable.) The ecological and other related fallacies are extensively discussed by Alker (1969), Diez-Roux (1998), and Blakely and Woodward (2000). King (1997), originally focusing on deriving correlates of individual voting behavior from aggregate data, describes a method for making inferences – within certain boundaries – at the micro level, when data are only available in aggregate form at the macro level.

The third potential error is the neglect of the original data structure, especially when some kind of analysis of covariance is to be used. Suppose one is interested in assessing between-school differences in pupil achievement after correcting for intake differences, and that Figure 3.1 depicts the true situation. The figure depicts the situation for five groups, for each of which we have five observations. The groups are indicated by the symbols \square , \times , $+$, \diamond , and $*$. The five group means are indicated by \bullet .

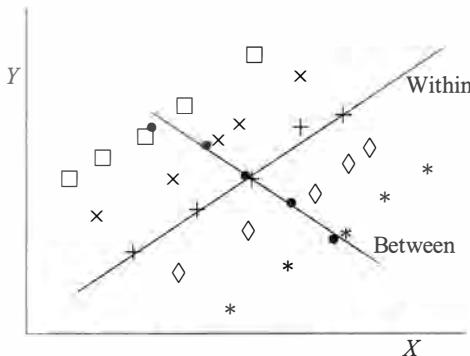


Figure 3.1: Micro-level versus macro-level adjustments.
 (X, Y) values for five groups indicated by $*, \diamond, +, \times, \square$; group averages by \bullet .

Now suppose the question is whether the differences between the groups on the variable Y , after adjusting for differences on the variable X , are substantial. The micro-level approach, which adjusts for the within-group regression of Y on X , will lead to the regression line with positive slope. In this picture, the micro-units from the group that have the \square

symbol are all above the line, whereas those from the group with the * symbol are all below the regression line. The micro-level regression approach will thus lead us to conclude that the five groups do differ given that an adjustment for X has been made.

Now suppose we were to aggregate the data, and regress the average \bar{Y} on the average \bar{X} . The averages are depicted by \bullet . This situation is represented in the graph by the regression line with negative slope. The averages of all groups are almost exactly on the regression line (the observed average \bar{Y} can be almost perfectly predicted from the observed average \bar{X}), thus leading us to the conclusion that there are almost no differences between the five groups after adjusting for the average \bar{X} .

Although the situation depicted in the graph is an idealized example, it clearly shows that working with aggregate data ‘is dangerous at best, and disastrous at worst’ (Aitkin and Longford, 1986, p. 42). When analyzing multilevel data, without aggregation, the problem described in this section can be dealt with by distinguishing between the within-group and the between-group regressions. This is worked out in Sections 3.6, 4.6, and 10.2.1.

The last objection to aggregation is that it prevents us from examining potential cross-level interaction effects of a specified micro-level variable with an as yet unspecified macro-level variable. Having aggregated the data to the macro level one cannot examine relations such as whether the sentence differential between black and white suspects is different between judges, when allowance is made for differences in seriousness of crimes. Or, to give another example, whether the effect of aptitude on achievement, present in the case of whole-class instruction, is smaller or even absent in the case of ability grouping of pupils within classrooms.

3.2 Disaggregation

Now suppose that we treat our data at the micro level. There are two situations:

1. we also have a measure of a variable at the macro level, next to the measures at the micro level;
2. we only have measures of micro-level variables.

In situation (1), disaggregation leads to ‘the miraculous multiplication of the number of units’. To illustrate, suppose a researcher is interested in the question of whether older judges hand down more lenient sentences than younger judges. A two-stage sample is taken: in the first stage ten judges are sampled, and in the second stage for each judge ten trials are sampled (in total there are thus $10 \times 10 = 100$ trials). One might disaggregate the data to the level of the trials and estimate the relation between the experience of the judge and the length of the sentence, without taking into account that some trials involve the same judge. This is like pretending that there are 100 independent observations, whereas in actual fact there are only 10 independent observations (the 10 judges). This shows that disaggregation and treating the data as if they are independent implies that the sample size is dramatically exaggerated. For the study of between-group differences, disaggregation often leads to serious risks of committing type I errors (asserting on the basis of the observations that there is a difference between older and younger judges whereas in the population of judges there is no such relation). On the other hand, when studying within-group differences, disaggregation often leads to unnecessarily conservative tests (i.e., type

I error probabilities that are too low); this is discussed in detail in Moerbeek et al. (2003) and Berkhof and Kampen (2004).

If measures are taken only at the micro level, analyzing the data at the micro level is a correct way to proceed, as long as one takes into account that observations within a macro-unit may be correlated. In sampling theory, this phenomenon is known as the design effect for two-stage samples. If one wants to estimate the average management capability of young managers, while in the first stage a limited number of organizations (say, 10) are selected and within each organization five managers are sampled, one runs the risk of making an error if (as is usually the case) there are systematic differences between organizations. In general, two-stage sampling leads to the situation that the ‘effective’ sample size that should be used to calculate standard errors is less than the total number of cases, the latter being given here by the 50 managers. The formula will be presented in one of the next sections.

Starting with Robinson’s (1950) paper on the ecological fallacy, many papers have been written about the possibilities and dangers of cross-level inference, that is, methods to conclude something about relations between micro-units on the basis of relations between data at the aggregate level, or conclude something about relations between macro-units on the basis of relations between disaggregated data. Discussions and many references are given by Pedhazur (1982, Chapter 13), Aitkin and Longford (1986), and Diez-Roux (1998). Our conclusion is that if the macro-units have any meaningful relation to the phenomenon under study, analyzing only aggregated or only disaggregated data is apt to lead to misleading and erroneous conclusions. A multilevel approach, in which within-group and between-group relations are combined, is more difficult but much more productive. This approach requires, however, assumptions to be specified about the way in which macro- and micro-effects are put together. The present chapter presents some multilevel procedures that are based on only a minimum of such assumptions (e.g., the additive model of equation (3.1)). Later chapters in this book are based on a more elaborate model, the so-called hierarchical linear model, which since about 1990 has been the most widely accepted basis for multilevel analysis.

3.3 The intraclass correlation

The degree of resemblance between micro-units belonging to the same macro-unit can be expressed by the *intraclass correlation coefficient*. The use of the term ‘class’ is conventional here and refers to the macro-units in the classification system under consideration. There are, however, several definitions of this coefficient, depending on the assumptions about the sampling design. In this section we assume a two-stage sampling design and infinite populations at either level. The macro-units will also be referred to as *groups*.

A relevant model here is the *random effects ANOVA* model.¹ Denoting by Y_{ij} the outcome value observed for micro-unit i within macro-unit j , this model can be expressed as

$$Y_{ij} = \mu + U_j + R_{ij}, \quad (3.1)$$

¹This model is also known in the statistical literature as the one-way random effects ANOVA model and as Eisenhart’s Type II ANOVA model. In multilevel modeling it is known as the empty model, and is treated further in Section 4.4.

where μ is the population grand mean, U_j is the specific effect of macro-unit j , and R_{ij} is the residual effect for micro-unit i within this macro-unit. In other words, macro-unit j has the ‘true mean’ $\mu + U_j$, and each measurement of a micro-unit within this macro-unit deviates from this true mean by some value R_{ij} . Units differ randomly from one another, which is reflected in the fact that U_j is a random variable and the name ‘random effects model’. Some units have a high true mean, corresponding to a high value of U_j , others have a true mean close to average, and still others a low true mean. It is assumed that all variables are independent, the group effects U_j having population mean 0 and population variance τ^2 (the *population between-group variance*), and the residuals having mean 0 and variance σ^2 (the *population within-group variance*). For example, if micro-units are pupils and macro-units are schools, then the within-group variance is the variance within the schools about their true means, while the between-group variance is the variance between the schools’ true means. The total variance of Y_{ij} is then equal to the sum of these two variances,

$$\text{var}(Y_{ij}) = \tau^2 + \sigma^2.$$

The number of micro-units within the j th macro-unit is denoted by n_j . The number of macro-units is N , and the total sample size is $M = \sum_j n_j$.

In this situation, the intraclass correlation coefficient ρ_I can be defined as

$$\rho_I = \frac{\text{population variance between macro-units}}{\text{total variance}} = \frac{\tau^2}{\tau^2 + \sigma^2}. \quad (3.2)$$

This is the proportion of variance that is accounted for by the group level. This parameter is called a correlation coefficient because it is equal to the correlation between values of two randomly drawn micro-units in the same, randomly drawn, macro-unit. Hedges and Hedberg (2007) report on a large variety of studies of educational performance in American schools, and find that values often range between 0.10 and 0.25.

It is important to note that the population variance between macro-units is not directly reflected by the observed variance between the means of the macro-units (the observed between-macro-unit variance). The reason is that in a two-stage sample, variation between micro-units will also show up as extra observed variance between macro-units. It is indicated below how the observed variance between cluster means must be adjusted to yield a good estimator for the population variance between macro-units.

Example 3.1 Random data.

Suppose we have a series of 100 observations as in the random digits in Table 3.1. The core part of the table contains the random digits. Now suppose that each row in the table is a macro-unit, so that for each macro-unit we have observations on 10 micro-units. The averages of the scores for each macro-unit are in the last column. There seem to be large differences between the randomly constructed macro-units, if we look at the variance in the macro-unit averages (which is 105.7). The total observed variance between the 100 micro-units is 814.0. Suppose the macro-units were schools, the micro-units pupils, and the random digits test scores. According to these two observed variances we might conclude that the schools differ considerably with respect to their average test scores. We know in this case, however, that in ‘reality’ the macro-units differ only by chance.

The following subsections show how the intraclass correlation can be estimated and tested. For a review of various inference procedures for the intraclass correlation we refer to Donner (1986). An extensive overview of many methods for estimating and testing the within-group and between-group variances is given by McCulloch and Searle (2001).

Table 3.1: Data grouped into macro-units (random digits from Glass and Stanley, 1970, p. 511).

j	Scores Y_{ij} for micro-units (random digits)										Average \bar{Y}_j
01	60	36	59	46	53	35	07	53	39	49	43.7
02	83	79	94	24	02	56	62	33	44	42	51.9
03	32	96	00	74	05	36	40	98	32	32	44.5
04	19	32	25	38	45	57	62	05	26	06	31.5
05	11	22	09	47	47	07	39	93	74	08	35.7
06	31	75	15	72	60	68	98	00	53	39	51.1
07	88	49	29	93	82	14	45	40	45	04	48.9
08	30	93	44	77	44	07	48	18	38	28	42.7
09	22	88	84	88	93	27	49	99	87	48	68.5
10	78	21	21	69	93	35	90	29	12	86	53.4

3.3.1 Within-group and between-group variance

We continue to refer to the macro-units as groups. To disentangle the information contained in the data about the population between-group variance and the population within-group variance, we consider the *observed variance between groups* and the *observed variance within groups*. These are defined as follows. The mean of macro-unit j is denoted by

$$\bar{Y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij}$$

and the overall mean is

$$\bar{Y}_{..} = \frac{1}{M} \sum_{j=1}^N \sum_{i=1}^{n_j} Y_{ij} = \frac{1}{M} \sum_{j=1}^N n_j \bar{Y}_j.$$

The observed variance within group j is given by

$$S_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2.$$

This number will vary from group to group. To have one parameter that expresses the within-group variability for all groups jointly, one uses the observed within-group variance, or pooled within-group variance. This is a weighted average of the variances within the various macro-units, defined as

$$\begin{aligned} S_{\text{within}}^2 &= \frac{1}{M - N} \sum_{j=1}^N \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2 \\ &= \frac{1}{M - N} \sum_{j=1}^N (n_j - 1) S_j^2. \end{aligned} \tag{3.3}$$

If model (3.1) holds, the expected value of the observed within-group variance is exactly equal to the population within-group variance:

$$\text{Expected variance within} = \mathcal{E}S_{\text{within}}^2 = \sigma^2. \quad (3.4)$$

The situation for the between-group variance is a little more complicated. For equal group sizes n_j , the observed between-group variance is defined as the variance between the group means,

$$S_{\text{between}}^2 = \frac{1}{N-1} \sum_{j=1}^N (\bar{Y}_j - \bar{Y}_{..})^2. \quad (3.5)$$

For unequal group sizes, the contributions of the various groups need to be weighted. The following formula uses weights that are useful for estimating the population between-group variance:

$$S_{\text{between}}^2 = \frac{1}{\tilde{n}(N-1)} \sum_{j=1}^N n_j (\bar{Y}_j - \bar{Y}_{..})^2. \quad (3.6)$$

In this formula, \tilde{n} is defined by

$$\tilde{n} = \frac{1}{N-1} \left\{ M - \frac{\sum_j n_j^2}{M} \right\} = \bar{n} - \frac{s^2(n_j)}{N\bar{n}}, \quad (3.7)$$

where $\bar{n} = M/N$ is the mean sample size and

$$s^2(n_j) = \frac{1}{N-1} \sum_{j=1}^N (n_j - \bar{n})^2$$

is the variance of the sample sizes. If all n_j have the same value, then \tilde{n} also has this value. In this case, S_{between}^2 is just the variance of the group means, given by (3.5).

It can be shown that the total observed variance is a combination of the within-group and the between-group variances, expressed as follows:

$$\begin{aligned} \text{observed total variance} &= \frac{1}{M-1} \sum_{j=1}^N \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{..})^2 \\ &= \frac{M-N}{M-1} S_{\text{within}}^2 + \frac{\tilde{n}(N-1)}{M-1} S_{\text{between}}^2. \end{aligned} \quad (3.8)$$

The complications with respect to the between-group variance arise from the fact that the micro-level residuals R_{ij} also contribute, albeit to a minor extent, to the observed between-group variance. Statistical theory tells us that the expected between-group variance is given by

$$\begin{aligned} \text{Expected observed variance between} \\ = \text{True variance between} + \text{Expected sampling error variance}. \end{aligned}$$

More specifically, the formula is

$$\mathcal{E}S_{\text{between}}^2 = \tau^2 + \frac{\sigma^2}{\tilde{n}} \quad (3.9)$$

(cf. Hays (1988, Section 13.3) for the case with constant n_j and Searle et al. (1992, Section 3.6) for the general case), which holds provided that model (3.1) is valid. The second term in this formula becomes small when \tilde{n} becomes large. Thus for large group sizes, the expected observed between variance is practically equal to the true between variance. For small group sizes, however, it tends to be larger than the true between variance due to the random differences that also exist between the group means.

In practice, we do not know the population values of the between and within macro-unit variances; these have to be estimated from the data. One way of estimating these parameters is based on formulas (3.4) and (3.9). From the former it follows that the population within-group variance, σ^2 , can be estimated without bias by the observed within-group variance:

$$\hat{\sigma}^2 = S_{\text{within}}^2. \quad (3.10)$$

From the combination of the last two formulas it follows that the population between-group variance, τ^2 , can be estimated without bias by taking the observed between-group variance and subtracting the contribution that the true within-group variance makes, on average, according to (3.9), to the observed between-group variance:

$$\hat{\tau}^2 = S_{\text{between}}^2 - \frac{S_{\text{within}}^2}{\tilde{n}}. \quad (3.11)$$

(Another expression is given in (3.14).) This expression can take negative values. This happens when the difference between group means is less than would be expected on the basis of the within-group variability, even if the true between-group variance τ^2 were 0. In such a case, it is natural to estimate τ^2 as 0.

It can be concluded that the split between the observed within-group variance and observed between-group variance does not correspond precisely to the split between the within-group and between-group variances in the population: the observed between-group variance reflects the population between-group variance plus a little of the population within-group variance.

The intraclass correlation is estimated according to formula (3.2) by

$$\hat{\rho}_I = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}^2}. \quad (3.12)$$

(Formula (3.15) gives another, equivalent, expression.) The standard error of this estimator in the case where all group sizes are constant, $n_j = n$, is given by

$$\text{S.E.}(\hat{\rho}_I) = (1 - \rho_I)(1 + (n - 1)\rho_I) \sqrt{\frac{2}{n(n - 1)(N - 1)}}. \quad (3.13)$$

This formula was given by Fisher (1958, Section 39) and by Donner (1986, equation (6.1)), who also gives the (quite complicated) formula for the standard error for the case of variable group sizes. Donner and Wells (1986) compare various ways to construct confidence intervals for the intraclass correlation coefficient.

The estimators given above are so-called analysis of variance or ANOVA estimators. They have the advantage that they can be represented by explicit formulas. Other much used estimators are those produced by the maximum likelihood (ML) and residual maximum likelihood (REML) methods (cf. Section 4.7). For equal group sizes, the ANOVA estimators are the same as the REML estimators (Searle et al., 1992). For unequal group sizes, the ML and REML estimators are slightly more efficient than the ANOVA estimators. Multilevel software can be used to calculate the ML and REML estimates.

Example 3.2 Within- and between-group variability for random data.

For our random digits table of the earlier example the observed between variance is $S_{\text{between}}^2 = 105.7$. The observed variance within the macro-units can be computed from formula (3.8). The observed total variance is known to be 814.0 and the observed between variance is given by 105.7. Solving (3.8) for the observed within variance yields $S_{\text{within}}^2 = (99/90) \times (814.0 - (10/11) \times 105.7) = 789.7$. Then the estimated true variance within the macro-units is also $\hat{\sigma}^2 = 789.7$. The estimate for the true between macro-units variance is computed from (3.11) as $\hat{\tau}^2 = 105.7 - (789.7/10) = 26.7$. Finally, the estimate of the intraclass correlation is $\hat{\rho}_I = 26.7/(789.7 + 26.7) = 0.03$. Its standard error, computed from (3.13), is 0.06.

3.3.2 Testing for group differences

The intraclass correlation as defined by (3.2) can be zero or positive.² A statistical test can be performed to investigate whether a positive value for this coefficient could be attributed to chance. If it may be assumed that the within-group deviations R_{ij} are normally distributed, one can use an exact test for the hypothesis that the intraclass correlation is 0, which is the same as the null hypothesis that there are no group differences, or the true between-group variance is 0. This is just the F -test for a group effect in the one-way analysis of variance, which can be found in any textbook on ANOVA. The test statistic can be written as

$$F = \frac{\tilde{n}S_{\text{between}}^2}{S_{\text{within}}^2},$$

and it has an F distribution with $N - 1$ and $M - N$ degrees of freedom if the null hypothesis holds.

Example 3.3 The F -test for the random data set.

For the data of Table 3.1, $F = (10 \times 105.7) / 789.7 = 1.34$ with 9 and 90 degrees of freedom. This value is far from significant ($p > 0.10$). Thus, there is no evidence of true between-group differences.

Statistical computer packages usually give the F -statistic and the within-group variance, S_{within}^2 . From this output, the estimated population between-group variance can be calculated by

$$\hat{\tau}^2 = \frac{S_{\text{within}}^2}{\tilde{n}}(F - 1) \tag{3.14}$$

²In a data set it is possible for the estimated intraclass correlation coefficient to be negative. This is always the case, for example, for group-centered variables. In a population satisfying model (3.1), however, the population intraclass correlation cannot be negative.

and the estimated intraclass correlation coefficient by

$$\hat{\rho}_I = \frac{F - 1}{F + \tilde{n} - 1}, \quad (3.15)$$

where \tilde{n} is given by (3.7). If $F < 1$, it is natural to replace both of these expressions by 0. These formulas show that a high value for the F -statistic will lead to large estimates for the between-group variance as well as the intraclass correlation, but that the group sizes, as expressed by \tilde{n} , moderate the relation between the test statistic and the parameter estimates.

If there are covariates, it often is relevant to test whether there are group differences in addition to those accounted for by the effect of the covariates. This is achieved by the usual F -test for the group effect in an analysis of covariance (ANCOVA). Such a test is relevant because it is possible that the ANOVA F -test does not demonstrate any group effects, but that such effects do emerge when controlling for the covariates (or vice versa). Another check on whether the groups make a difference can be carried out by testing the group-by-covariate interaction effect. These tests can be found in textbooks on ANOVA and ANCOVA, and they are contained in the well-known general-purpose statistical computer programs.

So, to test whether a given nesting structure in a data set calls for multilevel analysis, one can use standard ANOVA techniques. In addition to testing for the main group effect, it is also advisable to test for group-by-covariate interactions. If there is neither evidence for a main effect nor for interaction effects involving the group structure, then the researcher may leave aside the nesting structure and analyze the data by single-level methods such as ordinary least squares ('OLS') regression analysis. This approach to testing for group differences can be employed whenever the number of groups is not too large for the computer program being used. If there are too many groups, however, the program will refuse to do the job. In such a case it will still be possible to carry out the tests for group differences that are treated in the following chapters, following the logic of the hierarchical linear model. This will require the use of statistical multilevel software.

3.4 Design effects in two-stage samples

In the design of empirical investigations, the determination of sample sizes is an important decision. For two-stage samples, this is more complicated than for simple ('one-stage') random samples. An elaborate treatment of this question is given in Cochran (1977). This section gives a simple approach to the precision of estimating a population mean, indicating the basic role played by the intraclass correlation. We return to this question in Chapter 11.

Large samples are preferable in order to increase the precision of parameter estimates, that is, to obtain tight confidence intervals around the parameter estimates. In a simple random sample the standard error of the mean is related to the sample size by the formula

$$\text{standard error} = \frac{\text{standard deviation}}{\sqrt{\text{sample size}}}. \quad (3.16)$$

This formula can be used to indicate the required sample size (in a simple random sample) if a given standard error is desired.

When using two-stage samples, however, the clustering of the data should be taken into account when determining the sample size. Let us suppose that all group sizes are equal, $n_j = n$ for all j . Then the (total) sample size is Nn . The *design effect* is a number that indicates how much the sample size in the denominator of (3.16) is to be adjusted because of the sampling design used. It is the ratio of the variance of estimation obtained with the given sampling design to the variance of estimation obtained for a simple random sample from the same population, supposing that the total sample size is the same. A large design effect implies a relatively large variance, which is a disadvantage that may be offset by the cost reductions implied by the design. The design effect of a two-stage sample with equal group sizes is given by

$$\text{design effect} = 1 + (n - 1) \rho_I. \quad (3.17)$$

This formula expresses the fact that, from a purely statistical point of view, a two-stage sample becomes less attractive as ρ_I increases (clusters become more homogeneous) and as the group size n increases (the two-stage nature of the sampling design becomes stronger).

Suppose, for example, we were studying the satisfaction of patients with the treatments provided by their doctors. Furthermore, let us assume that some doctors have more satisfied patients than others, leading to a ρ_I of 0.30. The researchers used a two-stage sample, since that is far cheaper than selecting patients simply at random. They first randomly selected 100 doctors, from each chosen doctor selected five patients at random, and then interviewed each of these. In this case the design effect is $1 + (5 - 1) \times 0.30 = 2.20$. When estimating the standard error of the mean, we no longer can treat the observations as independent of each other. The effective sample size, that is, the equivalent total sample size that we should use in estimating the standard error, is equal to

$$N_{\text{effective}} = \frac{Nn}{\text{design effect}}, \quad (3.18)$$

in which N is the number of selected macro-units. For our example we find $N_{\text{effective}} = (100 \times 5) / 2.20 = 227$. So the two-stage sample with a total of 500 patients here is equivalent to a simple random sample of 227 patients.

One can also derive the total sample size using a two-stage sampling design on the basis of a desired level of precision, assuming that ρ_I is known, and fixing n because of budgetary or time-related considerations. The general rule is: this required sample size increases as ρ_I increases and it increases with the number of micro-units one wishes to select per macro-unit. Using (3.17) and (3.18), this can be derived numerically from the formula

$$N_{ts} = N_{srs} + N_{srs}(n - 1) \rho_I.$$

The quantity N_{ts} in this formula refers to the total desired sample size when using a two-stage sample, whereas N_{srs} refers to the desired sample size if one had used a simple random sample.

In practice, ρ_I is unknown. However, it often is possible to make an educated guess about it on the basis of earlier research.

In Figure 3.2, N_{ts} is graphed as a function of n and ρ_I (0.1, 0.2, 0.4, and 0.6, respectively), and taking $N_{srs} = 100$ as the desired sample size for an equally informative simple random sample.

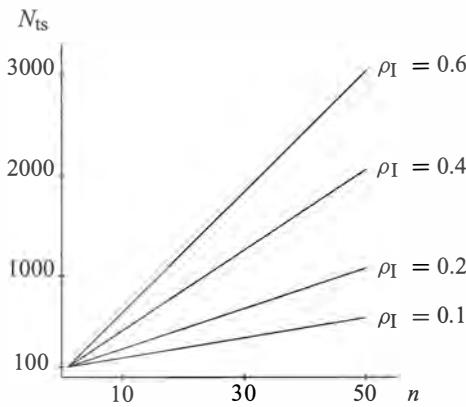


Figure 3.2: The total desired sample size in two-stage sampling.

3.5 Reliability of aggregated variables

Reliability, as conceived in psychological test theory (e.g., Lord and Novick, 1968) and in generalizability theory (e.g., Shavelson and Webb, 1991), is closely related to clustered data – although this may not be obvious at first sight. Classical psychological test theory considers a subject (an individual or other observational unit) with a given *true score*, of which imprecise, or unreliable, observations may be made. The observations can be considered to be nested within the subjects. If there is more than one observation per subject, the data are clustered. Whether there is only one observation or several, equation (3.1) is the model for this situation: the true score of subject j is $\mu + U_j$ and the i th observation on this subject is Y_{ij} , with associated measurement error R_{ij} . If several observations are taken, these can be aggregated to the mean value \bar{Y}_j which is then the measurement for the true score of subject j .

The same idea can be used when it is not an individual subject that is to be measured, but some collective entity: a school, a firm, or in general any macro-level unit such as those mentioned in Table 2.2. For example, when the school climate is measured on the basis of questions posed to pupils of the school, then Y_{ij} could refer to the answer by pupil i in school j to a given question, and the opinion of the pupils about this school would be measured by the mean value \bar{Y}_j . In terms of psychological test theory, the micro-level units i are regarded as parallel items for measuring the macro-level unit j .

The reliability of a measurement is defined generally as

$$\text{reliability} = \frac{\text{variance of true scores}}{\text{variance of observed scores}}.$$

It can be proved that this is equal to the correlation between independent replications of measuring the same subject. (In the mathematical model this means that the same value U_j

is measured, but with independent realizations of the random errors R_{ij} .) The reliability is indicated by the symbol λ_j .³

For measurement on the basis of a single observation according to model (3.1), the reliability is just the intraclass correlation coefficient:

$$\lambda_j = \rho_I = \frac{\tau^2}{\tau^2 + \sigma^2} \quad (\text{if } n_j = 1). \quad (3.19)$$

When several measurements are made for each macro-level unit, these constitute a cluster or group of measurements which are aggregated to the group mean \bar{Y}_j . To apply the general definition of reliability to \bar{Y}_j , note that the observed variance is the variance between the observed means \bar{Y}_j , while the true variance is the variance between the true scores $\mu + U_j$. Therefore the reliability of the aggregate is

$$\text{reliability of } \bar{Y}_j = \frac{\text{variance between } \mu + U_j}{\text{variance between } \bar{Y}_j}. \quad (3.20)$$

Example 3.4 Reliability for random data.

If in our previous random digits example the digits represented, for example, the perceptions by teachers in schools of their working conditions, then the aggregated variable, an indicator for organizational climate, has an estimated reliability of $26.7/105.7 = 0.25$. (The population value of this reliability is 0, however, as the data are random, so the true variance is 0.)

It can readily be demonstrated that the reliability of aggregated variables increases as the number of micro-units per macro-unit increases, since the true variance of the group mean (with group size n_j) is τ^2 while the expected observed variance of the group mean is $\tau^2 + \sigma^2/n_j$. Hence the reliability can be expressed by

$$\lambda_j = \frac{\tau^2}{\tau^2 + \sigma^2/n_j} = \frac{n_j \rho_I}{1 + (n_j - 1) \rho_I}. \quad (3.21)$$

It is quite clear that if n_j is very large and ρ_I is positive, then λ_j is almost 1. If $n_j = 1$ we are not able to distinguish between within- and between-group variance. Figure 3.3 presents a graph in which the reliability of an aggregate is depicted as a function of n_j (denoted by n) and ρ_I (0.1 and 0.4, respectively).

Much more can be said about constructing higher-level variables by aggregating lower-level variables; see, for example, Raudenbush and Sampson (1999b) and van Mierlo et al. (2009).

3.6 Within- and between-group relations

We saw in Section 3.1 that regressions at the macro level between aggregated variables \bar{X} and \bar{Y} can be completely different from the regressions between the micro-level variables X and Y . This section considers in more detail the interplay between macro-level and micro-level relations between two variables. First the focus is on regression of Y on X , then on the correlation between X and Y .

³In the literature the reliability of a measurement X is frequently denoted by the symbol ρ_{XX} , so that the reliability coefficient λ_j here could also be denoted by ρ_{YY} .

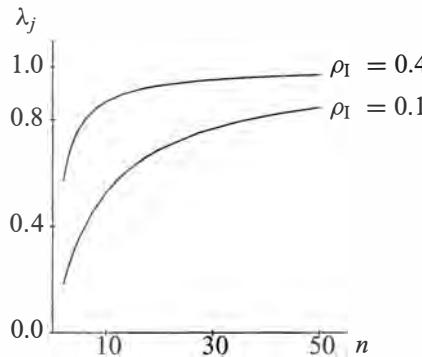


Figure 3.3: Reliability of aggregated variables.

The main point of this section is that within-group relations can be, in principle, completely different from between-group relations. This is natural, because the processes at work within groups may be different from the processes at work between groups (see Section 3.1). Total relations, that is, relations at the micro level when the clustering into macro-units is disregarded, are mostly a kind of average of the within-group and between-group relations. Therefore it is necessary to consider within- and between-group relations jointly, whenever the clustering of micro-units in macro-units is meaningful for the phenomenon being studied.

3.6.1 Regressions

The linear regression of a ‘dependent’ variable Y on an ‘explanatory’ or ‘independent’ variable X is the linear function of X that yields the best⁴ prediction of Y . When the bivariate distribution of (X, Y) is known and the data structure has only a single level, the expression for this regression function is

$$Y = \beta_0 + \beta_1 X + R,$$

where the regression coefficients are given by

$$\begin{aligned}\beta_0 &= \mathcal{E}(Y) - \beta_1 \mathcal{E}(X), \\ \beta_1 &= \frac{\text{cov}(X, Y)}{\text{var}(X)}.\end{aligned}$$

The constant term β_0 is called the *intercept*, while β_1 is called the regression coefficient. The term R is the *residual* or *error* component, and expresses the part of the dependent variable Y that cannot be approximated by a linear function of Y . Recall from Section 1.2.2 that $\mathcal{E}(X)$ and $\mathcal{E}(Y)$ denote the population mean (expected value) of X and Y , respectively.

⁴‘Best prediction’ means here the prediction that has the smallest mean squared error: the so-called least squares criterion.

Table 3.2: Artificial data on five macro-units, each with two micro-units.

j	i	X_{ij}	\bar{X}_j	Y_{ij}	\bar{Y}_j
1	1	1	2	5	6
1	2	3	2	7	6
2	1	2	3	4	5
2	2	4	3	6	5
3	1	3	4	3	4
3	2	5	4	5	4
4	1	4	5	2	3
4	2	6	5	4	3
5	1	5	6	1	2
5	2	7	6	3	2

In a multilevel data structure, this principle can be applied in various ways, depending on which population of X and Y values is being considered.

Let us consider the artificial data set of Table 3.2. The first two columns in the table contain the identification numbers of the macro-unit (j) and the micro-unit (i). The other four columns contain the data. X_{ij} is the variable observed for micro-unit i in macro-unit j , \bar{X}_j the average of the X_{ij} values for group j , and similarly for the dependent variable Y .

One might be interested in the relation between Y_{ij} and X_{ij} . The linear regression of Y_{ij} on X_{ij} at the micro level for the total group of 10 observations is

$$Y_{ij} = 5.33 - 0.33 X_{ij} + R. \quad (\text{total regression})$$

This is the disaggregated relation, since the nesting of micro-units in macro-units is not taken into account. The regression coefficient is -0.33 .

The aggregated relation is the linear regression relationship at the macro level of the group means \bar{Y}_j on the group means \bar{X}_j . This regression line is

$$\bar{Y}_j = 8.00 - 1.00 \bar{X}_j + R. \quad (\text{regression between group means})$$

The regression coefficient is now -1.00 .

A third option is to describe the relation between Y_{ij} and X_{ij} within each single group. Assuming that the regression coefficient has the same value in each group, this is the same as the regression of the within-group Y -deviations ($Y_{ij} - \bar{Y}_j$) on the X -deviations ($X_{ij} - \bar{X}_j$). This within-group regression line is given by

$$Y_{ij} = \bar{Y}_j + 1.00 (X_{ij} - \bar{X}_j) + R, \quad (\text{regression within groups})$$

with a regression coefficient of $+1.00$.

Finally (and this is how the artificial data set was constructed), Y_{ij} can be written as a function of the within-group and between-group relations between Y and X . This amounts to putting together the between-group and within-group regression equations. The result is

$$\begin{aligned} Y_{ij} &= 8.00 - 1.00 \bar{X}_j + 1.00 (X_{ij} - \bar{X}_j) + R \\ &= 8.00 + 1.00 X_{ij} - 2.00 \bar{X}_j + R. \end{aligned} \quad (\text{multilevel regression})$$

Figure 3.4 graphically depicts the total, within-group, and between-group relations between the variables.

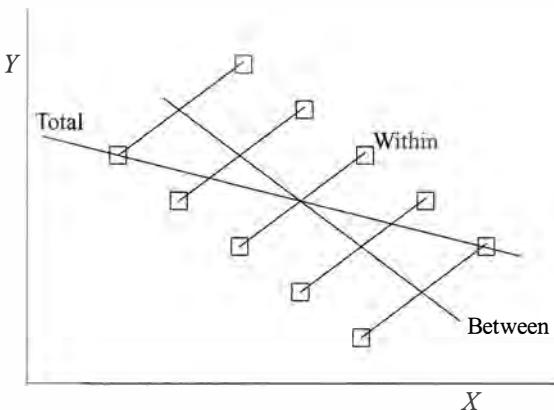


Figure 3.4: Within, between, and total relations.

The five parallel ascending lines represent the within-group relation between Y and X . The steep descending line represents the relation at the aggregate level (i.e., between the group means), whereas the almost horizontal descending line represents the total relationship, that is, the micro-level relation between X and Y ignoring the hierarchical structure.

The within-group regression coefficient is +1 whereas the between-group coefficient is -1. The total regression coefficient, -0.33, lies in between these two. This illustrates that within-group and between-group relations can be completely different, even have opposite signs. The true relation between Y and X is revealed only when the within- and between-group relations are considered jointly, that is, by the multilevel regression. In the multilevel regression, both the between-group and within-group regression coefficients play a role. Thus there are many different ways to describe the data, of which one is the best, because it describes how the data were generated. In this artificial data set, the residual R is 0; real data, of course, have nonzero residuals.

A population model

The interplay of within-group and between-group relations can be better understood on the basis of a population model such as (3.1). Since this section is about two variables, X and Y , a bivariate version of the model is needed. In this model, group (macro-unit) j has specific main effects U_{xj} and U_{yj} for variables X and Y , and associated with individual (micro-unit) i are the within-group deviations R_{xij} and R_{yij} . The population means are denoted by μ_x and μ_y and it is assumed that the Us and the Rs have population means 0. The Us on the one hand and the Rs on the other are independent. The formulas for X and Y are then

$$\begin{aligned} X_{ij} &= \mu_x + U_{xj} + R_{xij}, \\ Y_{ij} &= \mu_y + U_{yj} + R_{yij}. \end{aligned} \tag{3.22}$$

For the formulas that refer to relations between group means \bar{X}_j and \bar{Y}_j , it is assumed that each group has the same size, denoted by n .

The correlation between the group effects is defined as

$$\rho_{\text{between}} = \rho(U_{xj}, U_{yj}),$$

while the correlation between the individual deviations is defined by

$$\rho_{\text{within}} = \rho(R_{xij}, R_{yij}).$$

One of the two variables X and Y might have a stronger group nature than the other, so that the intraclass correlation coefficients for X and Y may be different. These are denoted by ρ_{lx} and ρ_{ly} , respectively.

The *within-group regression coefficient* is the regression coefficient within each group of Y on X , assumed to be the same for each group. This coefficient is denoted by β_{within} and defined by the within-group regression equation,

$$Y_{ij} = \mu_y + U_{yj} + \beta_{\text{within}}(X_{ij} - \mu_x - U_{xj}) + R. \quad (3.23)$$

This equation may be regarded as an analysis of covariance model for Y . Hence the within-group regression coefficient is also the effect of X in the ANCOVA approach to this multilevel data.

The within-group regression coefficient is also obtained when the Y -deviation values ($Y_{ij} - \bar{Y}_j$) are regressed on the X -deviation values ($X_{ij} - \bar{X}_j$). In other words, it is also the regression coefficient obtained in the disaggregated analysis of the within-group deviation scores.

The population *between-group regression coefficient* is defined as the regression coefficient for the group effects U_y on U_x . This coefficient is denoted by $\beta_{\text{between } U}$ and is defined by the regression equation

$$U_{yj} = \beta_{\text{between } U} U_{xj} + R,$$

where R now is the group-level residual.

The *total regression coefficient* of Y on X is the regression coefficient in the disaggregated analysis, that is, when the data are treated as single-level data:

$$Y_{ij} = \mu_y + \beta_{\text{total}}(X_{ij} - \mu_x) + R.$$

The total regression coefficient can be expressed as a weighted mean of the within-and between-group coefficients, where the weight for the between-group coefficient is just the intraclass correlation for X . The formula is

$$\beta_{\text{total}} = \rho_{lx} \beta_{\text{between } U} + (1 - \rho_{lx}) \beta_{\text{within}}. \quad (3.24)$$

This expression implies that if X is a pure macro-level variable (so that $\rho_{lx} = 1$), the total regression coefficient is equal to the between-group coefficient. Conversely, if X is a pure micro-level variable we have $\rho_{lx} = 0$, and the total regression coefficient is just the within-group coefficient. Usually X will have both a within-group and a between-group component and the total regression coefficient will be somewhere between the two level-specific regression coefficients.

At the macro level, the regression of the *observed* group means \bar{Y}_j on \bar{X}_j is not the same as the regression of the ‘true’ group effects U_j on U_x . This is because the observed group averages, \bar{X}_j and \bar{Y}_j , can be regarded as the ‘true’ group means to which some error, $\bar{R}_{x,j}$ and $\bar{R}_{y,j}$, has been added.⁶ Therefore the regression coefficient for the observed group means is not exactly equal to the (population) between-group regression coefficient, but it is given by

$$\beta_{\text{between group means}} = \lambda_{xj} \beta_{\text{between } U} + (1 - \lambda_{xj}) \beta_{\text{within}}, \quad (3.25)$$

where λ_{xj} is the reliability of the group means \bar{X}_j for measuring $\mu_x + U_{xj}$, given by equation (3.21) applied to the X -variable. If n is large the reliability will be close to unity, and the regression coefficient for the group means will be close to the between-group regression coefficient at the population level.

Combining equations (3.24) and (3.25) leads to another expression for the total regression coefficient. This expression uses the correlation ratio η_x^2 which is defined as the ratio of the intraclass correlation coefficient to the reliability of the group mean,

$$\eta_x^2 = \frac{\rho_{Ix}}{\lambda_{xj}} = \frac{\tau_x^2 + \sigma_x^2/n}{\tau_x^2 + \sigma_x^2} = \rho_{Ix} + \frac{1}{n}(1 - \rho_{Ix}). \quad (3.26)$$

For large group sizes the reliability approaches unity, so the correlation ratio approaches the intraclass correlation.

In the data, the correlation ratio η_x^2 is the same as the proportion of variance in X_{ij} explained by the group means, and it can be computed as the ratio of the between-group sum of squares relative to the total sum of squares in an analysis of variance, that is,

$$\hat{\eta}_x^2 = \frac{\sum_j n_j (\bar{X}_j - \bar{X}_{..})^2}{\sum_{i,j} (X_{ij} - \bar{X}_{..})^2}.$$

The combined expression indicates how the total regression coefficient depends on the within-group regression coefficient and the regression coefficient between the group means:

$$\beta_{\text{total}} = \eta_x^2 \beta_{\text{between group means}} + (1 - \eta_x^2) \beta_{\text{within}}. \quad (3.27)$$

Expression (3.27) was first given by Duncan et al. (1961) and can also be found, for example, in Pedhazur (1982, p. 538). A multivariate version was given by Maddala (1971). To apply this equation to an unbalanced data set, the regression coefficient between group means must be calculated in a weighted regression, group j having weight n_j .

Example 3.5 Within- and between-group regressions for artificial data.

In the artificial example given, the total sum of squares of X_{ij} as well as Y_{ij} is 30 and the between-group sums of squares for X and Y are 20. Hence the correlation ratios are $\hat{\eta}_x^2 = \hat{\eta}_y^2 = 20/30 = 0.667$. If we use this value and plug it into formula (3.27), we find

$$\hat{\beta}_{\text{total}} = 0.667 \times (-1.00) + (1 - 0.667) \times 1.00 = -0.33,$$

which is indeed what we found earlier.

⁵The remainder of this subsection may be skipped by the cursory reader.

⁶The same phenomenon is at the heart of formulas (3.9) and (3.11).

3.6.2 Correlations

The quite extreme nature of the artificial data set of Table 3.2 becomes apparent when we consider the correlations.

The group means (\bar{X}_j, \bar{Y}_j) lie on a decreasing straight line, so the *observed between-group correlation*, which is defined as the correlation between the group means, is $R_{\text{between}} = -1$. The *within-group correlation* is defined as the correlation within the groups, assuming that this correlation is the same within each group. This can be calculated as the correlation coefficient between the within-group deviation scores $\tilde{X}_{ij} = (X_{ij} - \bar{X}_j)$ and $\tilde{Y}_{ij} = (Y_{ij} - \bar{Y}_j)$. In this data set the deviation scores $(\tilde{X}_{ij}, \tilde{Y}_{ij})$ are $(-1, -1)$ for $i = 1$ and $(+1, +1)$ for $i = 2$, so the within-group correlation here is $R_{\text{within}} = +1$. Thus, we see that the within-group as well as the between-group correlations are perfect, but of opposite signs. The disaggregated correlation, that is, the correlation computed without taking the nesting structure into account, is $R_{\text{total}} = -0.33$. (This is the same as the value for the regression coefficient in the total (disaggregated) regression equation, because X and Y have the same variance.)

The population model again

Recall that in the population model mentioned above, the correlation coefficient between the group effects U_x and U_y was defined as ρ_{between} and the correlation between the individual deviations R_x and R_y was defined as ρ_{within} . The intraclass correlation coefficients for X and Y were denoted by ρ_{lx} and ρ_{ly} .

How do these correlations between unobservable variables relate to correlations between observables? The population within-group correlation is also the correlation between the within-group deviation scores $(\tilde{X}_{ij}, \tilde{Y}_{ij})$:

$$\rho(\tilde{X}_{ij}, \tilde{Y}_{ij}) = \rho_{\text{within}}. \quad (3.28)$$

For the between-group coefficient the relation is, as always, a little more complicated. The correlation coefficient between the group means is equal to

$$\rho(\bar{X}_j, \bar{Y}_j) = \sqrt{\lambda_{xj} \lambda_{yj}} \rho_{\text{between}} + \sqrt{(1 - \lambda_{xj})(1 - \lambda_{yj})} \rho_{\text{within}}, \quad (3.29)$$

where λ_{xj} and λ_{yj} are the reliability coefficients of the group means (see equations (3.20, 3.21)). For large group sizes the reliabilities will be close to 1 (provided the intraclass correlations are larger than 0), so that the correlation between the group means will then be close to ρ_{between} .

The total correlation (i.e., the correlation in the disaggregated analysis) is a combination of the within-group and between-group correlation coefficients. The combination depends on the intraclass correlations, as shown by the formula

$$\rho(X_{ij}, Y_{ij}) = \sqrt{\rho_{lx} \rho_{ly}} \rho_{\text{between}} + \sqrt{(1 - \rho_{lx})(1 - \rho_{ly})} \rho_{\text{within}}. \quad (3.30)$$

If the intraclass correlations are low, then X and Y primarily have the nature of level-one variables, and the total correlation will be close to the within-group correlation; on the other hand, if the intraclass correlations are close to 1, then X and Y almost have the nature

of level-two variables and the total correlation is close to the between-group correlation. As a third possibility, if one of the intraclass correlations is close to 0 and the other is close to 1, then one variable is mainly a level-one variable and the other mainly a level-two variable. Formula (3.30) then implies that the total correlation coefficient is close to 0, no matter how large the within-group and between-group correlations. This is intuitively obvious, since a level-one variable with hardly any between-group variability cannot be substantially correlated with a variable having hardly any within-group variability.

If the intraclass correlations of X and Y are equal and denoted by ρ_I , then (3.30) can be formulated more simply as

$$\rho(X_{ij}, Y_{ij}) = \rho_I \rho_{\text{between}} + (1 - \rho_I) \rho_{\text{within}}.$$

In this case the weights ρ_I and $(1 - \rho_I)$ add up to 1 and the total correlation coefficient is necessarily between the within-group and the between-group correlation coefficients. This is not true in general, because the sum of the weights in (3.30) is smaller than 1 if the intraclass correlations for X and Y are different.

The reliabilities of the group means then also are equal, $\lambda_{xj} = \lambda_{yj}$; let us denote these by λ_j . The correlation coefficient between the group means (3.29) then simplifies to

$$\rho(\bar{X}_j, \bar{Y}_j) = \lambda_j \rho_{\text{between}} + (1 - \lambda_j) \rho_{\text{within}}.$$

The last two formulae can help in understanding how aggregation changes correlations – under the continued assumption that the intraclass correlations of X and Y are the same. The total correlation as well as the correlation between group means are between the within-group and the between-group correlations. However, since the reliability coefficient λ_j is greater⁷ than the intraclass correlation ρ_I , the correlation between the group means is drawn toward the between-group correlation more strongly than the total correlation is. Therefore, aggregation will increase correlation *if and only if* the between-group correlation coefficient is larger than the within-group correlation coefficient. Therefore, the fact that correlations between group means are often higher than correlations between individuals is not the mathematical consequence of aggregation, but the consequence of the processes at the group level (determining the value of ρ_{between}) being different from the processes at the individual level (which determine the value of ρ_{within}).

Correlations between observed group means⁸

Analogous to the regression coefficients, for the correlation coefficients we can also combine the equations to see how the total correlation depends on the within-group correlation and the correlation between the group means. This yields

$$\rho(X_{ij}, Y_{ij}) = \eta_x \eta_y \rho(\bar{X}_j, \bar{Y}_j) + \sqrt{(1 - \eta_x^2)(1 - \eta_y^2)} \rho_{\text{within}}. \quad (3.31)$$

This expression was given by Robinson (1950) and can also be found, for example, in Pedhazur (1982, p. 536). When it is applied to an unbalanced data set, the correlation between the group means should be calculated with weights n_j .

⁷We argue under the assumption that the group size is at least 2, and the common intraclass correlation is positive.

⁸The remainder of Section 3.6.2 may also be skipped by the cursory reader.

It may be noted that many texts do not make the explicit distinction between population and data. If the population and the data are equated, then the reliabilities are unity, the correlation ratios are the same as the intraclass correlations, and the population between-group correlation is equal to the correlation between the group means. The equation for the total correlation then becomes

$$R_{\text{total}} = \hat{\eta}_x \hat{\eta}_y R_{\text{between}} + \sqrt{(1 - \hat{\eta}_x^2)(1 - \hat{\eta}_y^2)} R_{\text{within}}. \quad (3.32)$$

When parameter estimation is being considered, however, confusion may be caused by neglecting this distinction.

Example 3.6 Within- and between-group correlations for artificial data.

The correlation ratios in the artificial data example are $\eta_x^2 = \eta_y^2 = 0.667$ and we also saw above that $R_{\text{within}} = +1$ and $R_{\text{between}} = -1$. Filling in these numbers in formula (3.32) yields

$$R_{\text{total}} = \sqrt{0.667^2} \times (-1.00) + \sqrt{(1 - 0.667)^2} \times 1.00 = -0.33,$$

which indeed is the value found earlier for the total correlation.

3.6.3 Estimation of within- and between-group correlations

There are several ways of obtaining estimates for the correlation parameters treated in this section.

A quick method is based on the intraclass correlations, estimated as in Section 3.3.1 or from the output of a multilevel computer program, and the observed within-group and total correlations. The observed within-group correlation is just the ordinary correlation coefficient between the within-group deviations ($X_{ij} - \bar{X}_j$) and ($Y_{ij} - \bar{Y}_j$), and the total correlation is the ordinary correlation coefficient between X and Y in the whole data set. The quick method is then based on (3.28) and (3.30). This leads to the estimates

$$\hat{\rho}_{\text{within}} = R_{\text{within}} \quad (3.33)$$

and

$$\hat{\rho}_{\text{between}} = \frac{R_{\text{total}} - \sqrt{(1 - \hat{\rho}_{\text{lx}})(1 - \hat{\rho}_{\text{ly}})} R_{\text{within}}}{\sqrt{\hat{\rho}_{\text{lx}} \hat{\rho}_{\text{ly}}}}. \quad (3.34)$$

This is not statistically the most efficient method, but it is straightforward and leads to good results if sample sizes are not too small.

The ANOVA method (Searle, 1956) goes via the variances and covariances, based on the definition

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \text{var}(Y)}}.$$

Estimating the within- and between-group variances was discussed in Section 3.3.1. The within- and between-group covariances between X and Y can be estimated by formulas

analogous to (3.3), (3.6), (3.10) and (3.11), replacing the squares $(Y_{ij} - \bar{Y}_j)^2$ and $(\bar{Y}_j - \bar{Y}_{..})^2$ by the cross-products $(X_{ij} - \bar{X}_j)(Y_{ij} - \bar{Y}_j)$ and $(\bar{X}_j - \bar{X}_{..})(\bar{Y}_j - \bar{Y}_{..})$. It is shown in Searle et al. (1992, Section 11.1.a) how these calculations can be replaced by a calculation involving only sums of squares.

Finally, the maximum likelihood and residual maximum likelihood methods can be used. These are the estimation methods most often used (cf. Section 4.7) and are implemented in multilevel software. Chapter 16 describes multivariate multilevel models; the correlation coefficient between two variables refers to the simplest multivariate situation, namely, bivariate data. Formula (16.5) represents the model which allows the ML and REML estimation of within-group and between-group correlations.

Example 3.7 Within- and between-group correlations for school tests.

Many examples in Chapters 4 and 5 use a data set on the school performance of 3,758 pupils in $N = 211$ schools. The present example considers the correlation between the scores on an arithmetic test (X) and a language test (Y). We delete only pupils who have missing data on either of these variables, leaving $M = 3,762$ pupils. Within-school correlations reflect the correspondence between the pupils' language and arithmetic capabilities (attenuated by the unreliability of the tests). Between-school correlations reflect the processes that determine the schools' populations (intake policy, social segregation of neighborhoods) and the influence of teachers and school policy. Thus the within-school correlations and the between-school correlations are caused by quite different processes.

The indices 'within', 'between', and 'total' will be abbreviated to 'w', 'b', and 't'. The observed correlations are $R_w = 0.634$, $R_b = 0.871$, and $R_t = 0.699$.

The ANOVA estimators are calculated in accordance with the principles of Section 3.3.1. The observed within-group variances and covariance, which are also the estimated within-group population variances and covariance (cf. (3.10)), are $\hat{\sigma}_x^2 = S_{wx}^2 = 32.118$, $\hat{\sigma}_y^2 = S_{wy}^2 = 62.607$, $\hat{\sigma}_{xy} = S_{wxy} = 28.432$.

The observed between-group variances and covariance are $S_{bx}^2 = 13.405$, $S_{by}^2 = 19.966$, $S_{bxy} = 14.254$. For this data set, $\tilde{n} = 17.816$. According to (3.11), the estimated population between-group variances and covariance are $\hat{\tau}_x^2 = 11.602$, $\hat{\tau}_y^2 = 16.452$, $\hat{\tau}_{xy} = 12.658$.

From these estimated variances, the intraclass correlations are computed as $\rho_{Ix} = 11.602 / (11.602 + 32.118) = 0.2654$ and $\rho_{Iy} = 16.452 / (16.452 + 62.607) = 0.2081$.

The 'quick' method uses the observed within and total correlations and the intraclass correlations. The resulting estimates are $\hat{\rho}_w = 0.634$ from (3.33) and $\hat{\rho}_b = 0.915$ from (3.34).

The ANOVA estimates for the within- and between-group correlations use the estimated within- and between-group population variances and covariance. The results are $\hat{\rho}_w = 28.432 / \sqrt{32.118 \times 62.607} = 0.634$ and $\hat{\rho}_b = 12.658 / \sqrt{11.602 \times 16.452} = 0.916$.

The ML estimates for the within-group variances and covariance are obtained in Chapter 16 as $\hat{\sigma}_x^2 = 32.12$, $\hat{\sigma}_y^2 = 62.87$, $\hat{\sigma}_{xy} = 28.51$. For the between-group variances and covariance they are $\hat{\tau}_x^2 = 12.45$, $\hat{\tau}_y^2 = 17.86$, $\hat{\tau}_{xy} = 13.77$. This leads to estimated correlations $\hat{\rho}_w = 28.51 / \sqrt{32.12 \times 62.87} = 0.634$ and $\hat{\rho}_b = 13.77 / \sqrt{14576 \times 17.86} = 0.923$.

It can be concluded that, for this large data set, the three methods all yield practically the same results. The within-school (pupil-level) correlation, 0.63, is substantial. Thus pupils' language and arithmetic capacities are closely correlated. The between-school correlation, 0.92, is very high. This demonstrates that the school policies, the teaching quality, and the processes that determine the composition of the school population have practically the same effect on the pupils' language as on their arithmetic performance. Note that the observed between-school correlation, 0.87, is not quite as high because of the attenuation caused by unreliability that follows from (3.29).

3.7 Combination of within-group evidence

When research focuses on within-group relations and several groups (macro-units) have been investigated, it is often desired to combine the evidence gathered in the various groups. For example, consider a study in which a relation between work satisfaction (X) and sickness leave (Y) is studied in several organizations. If the organizations are sufficiently similar, or if they can be considered as a sample from some population of organizations, then the data can be analyzed according to the hierarchical linear model of the next chapters. This also requires that the number of organizations is large enough (e.g., 20 or more), because this number is the sample size from the population of organizations. On the other hand, if the organizations are too diverse and not representative of any population, or if there are not enough of them, there are still inferential procedures to combine the results across the organizations and provide a single answer on the relation between X and Y in which the evidence from all organizations is combined.

Another example is meta-analysis, the statistically based combination of several studies. There are many texts about meta-analysis, among them Hedges and Olkin (1985), Rosenthal (1991), Hedges (1992), Wilson and Lipsey (2001), and Bohrenstein et al. (2009). A number of publications may contain information about the same phenomenon, and it may be important to combine this information in a single test.

There exist various methods for combining evidence from several studies that are based only on the assumption that this evidence is statistically independent. They can be applied if the number of independent studies is at least two. The least demanding method is Fisher's combination of p -values (Fisher, 1958; Hedges and Olkin, 1985). This method assumes that in each of the N studies a null hypothesis is tested, which results in independent p -values p_1, \dots, p_N . The combined null hypothesis is that in *all* of the studies the null hypothesis holds; the combined alternative hypothesis is that in *at least* one of the studies the alternative hypothesis holds. It is not required that the N independent studies used the same operationalizations or methods of analysis, only that it is meaningful to test this combined null hypothesis. This hypothesis can be tested by minus twice the sum of the natural logarithms of the p -values,

$$\chi^2 = -2 \sum_{j=1}^N \ln(p_j), \quad (3.35)$$

which under the combined null hypothesis has a chi-squared distribution with $2N$ degrees of freedom. Because of the shape of the logarithmic function, this combined statistic will already have a large value if at least one of the p -values is very small.

A stronger combination procedure can be achieved if the several studies all lead to estimates of theoretically the same parameter, denoted here by θ . Suppose that the j th study yields a parameter estimate $\hat{\theta}_j$ with standard error s_j and that all the studies are statistically independent. Then the combined estimate with smallest standard error is the weighted average with weights inversely proportional to s_j^2 ,

$$\hat{\theta} = \frac{\sum_j s_j^{-2} \hat{\theta}_j}{\sum_j s_j^{-2}}, \quad (3.36)$$

with standard error

$$\text{S.E.}(\hat{\theta}) = \sqrt{\frac{1}{\sum_j s_j^{-2}}}. \quad (3.37)$$

For example, if standard errors are inversely proportional to the square root of sample size, $s_j = \sigma / \sqrt{n_j}$ for some value σ , then the weights are directly proportional to the sample sizes and the standard error of the combined estimate is $\sigma / \sqrt{\sum n_j}$. If the individual estimates are approximately normally distributed, or N is large even while the estimates are not nearly normally distributed, the t -ratio,

$$\frac{\hat{\theta}}{\text{S.E.}(\hat{\theta})},$$

can be tested in a standard normal distribution.

A third, intermediate combination procedure is based on the assumption that the parameters estimated by the several studies are not necessarily the same, but are independent random draws from the same population. If the true value of the parameter in population j is denoted by θ_j , this means that

$$\theta_j = \theta + E_j, \quad (3.38)$$

where θ is the mean parameter in the population of all potential studies, and E_j is the deviation from this value in study j , dependent on particulars of the group under study, the measurement instrument used, etc. The estimate can be represented as

$$\hat{\theta}_j = \theta_j + R_j,$$

where R_j is a random residual. From each study we know the standard error $s_j = \text{S.E.}(\hat{\theta}_j) = \sqrt{\text{var}(R_j)}$. Combining these two equations leads to a representation of the parameter estimates as a mean value plus a double error term,

$$\hat{\theta}_j = \theta + E_j + R_j. \quad (3.39)$$

If it is reasonable to assume that the studies are a random draw from a population, then the residuals E_j may be regarded as being independent with expected value 0 and a common unknown variance σ^2 . If each of the studies is large enough and the errors for each study are approximately normally distributed, then the residuals R_j can be regarded as normally distributed variables with expected value 0 and known variance s_j^2 . Then (3.39) represents a heteroscedastic model, with variances $\sigma^2 + s_j^2$. If the number of studies is large enough, then study-level variables may be added, giving, for one explanatory variable, a regression model such as

$$\hat{\theta}_j = \beta_0 + \beta_1 x_j + E_j + R_j. \quad (3.40)$$

Such two-stage models can be estimated with software such as HLM, MLwiN, or R. This model, and extended versions, are treated for meta-analysis, for example, by Raudenbush

and Bryk (2002, Chapter 7), where they are called the ‘V-known problem’. The model can be regarded as a special case of the hierarchical linear model treated in the following chapters, and as a two-step approach to multilevel analysis. In contrast to some earlier two-step approaches, it is correct in the sense that it recognizes the distinct contributions of within-study ‘error’ variability and between-study ‘true’ variability, a distinction that has permeated this chapter since Section 3.3.1. Achen (2005) discussed the suitability of this model for multilevel analysis, focusing on applications in political science.

The choice between these three combination methods can be made as follows. The combination of estimates, expressed by (3.36), is more suitable if conceptually it can be argued that the parameter being estimated in each of the combined studies does indeed have the same theoretical interpretation and it is plausible that this parameter (the estimated θ) has the same value in each of the studies. Models (3.39) and (3.40) with a double error term are applicable if it can be argued that the theoretical interpretation is the same, but the value of this parameter could differ across studies. On the other hand, Fisher’s method (3.35) for combining p -values requires only that the null hypotheses of ‘no effect’ can be regarded as the same, and it has good power even if the effect sizes are very different between the N studies. For example, the individual tests could be two-sided tests and the effects could be of different signs; as long as it is meaningful to test the combined null hypothesis that the null hypothesis holds in *all* individual studies, this method may be applied. More combination methods can be found in the literature about meta-analysis (e.g., Hedges and Olkin, 1985).

Sometimes it may be helpful to apply a formal statistical test of the null hypothesis that all the estimates $\hat{\theta}_j$ do indeed have the same expected value:

$$H_0: \mathcal{E}(\hat{\theta}_1) = \mathcal{E}(\hat{\theta}_2) = \dots = \mathcal{E}(\hat{\theta}_N), \text{ or equivalently } \theta_1 = \theta_2 = \dots = \theta_N. \quad (3.41)$$

Assuming that the standard errors were estimated very precisely, this can be done by a chi-squared test as derived, for example, in Lehmann and Romano (2005, Section 7.3). The test statistic is

$$C = \sum_j \left(\frac{\hat{\theta}_j - \hat{\theta}}{s_j} \right)^2 \quad (3.42)$$

and has, under the null hypothesis and if the estimates $\hat{\theta}_j$ are normally distributed with variances s_j^2 , a chi-squared distribution with $N - 1$ degrees of freedom.

Example 3.8 Gossip behavior in six organizations.

Wittek and Wielers (1998) investigated effects of informal social network structures on gossip behavior in six work organizations. One of the hypotheses tested was that individuals tend to gossip more if they are involved in more coalition triads. An individual A is involved in a coalition triad with two others, B and C , if he has a positive relation with B while A and B both have a negative relation with C . Six organizations were studied which were so different that an approach following the lines of the hierarchical linear model was not considered appropriate. For each organization separately, a multiple regression was carried out to estimate the effect of the number of coalition triads in which a person was involved on a measure for gossip behavior, controlling for some relevant other variables.

The p -values obtained were 0.015, 0.42, 0.19, 0.13, 0.25, and 0.43. Only one of these is significant (i.e., less than 0.05), and the question is whether this combination of six p -values would be unlikely under the combined null hypothesis which states that in *all* six organizations the effect of coalition triads on gossip is absent. Equation (3.35) yields the test statistic $\chi^2 = 22.00$ with

$2 \times 6 = 12$ degrees of freedom, $p < 0.05$. Thus the result is significant, which shows that in the combined data there is indeed evidence of an effect of coalition triads on gossip behavior, even though this effect is significant in only one of the organizations considered separately.

3.8 Glommary

Aggregation and disaggregation. Multilevel data structures can be analyzed by aggregating data to the higher level (e.g., by taking the means) and analyzing these; or by disaggregating to the lower level, which means that characteristics of higher-level units are used as characteristics of the lower-level units contained in them, and further nesting is not used. Both approaches provide only a limited perspective because they focus on only one level among several.

Shift of meaning. Variables aggregated from a lower to a higher level (e.g., the average income in a neighborhood or the proportions of girls in a classroom) have a theoretically different meaning at the level of aggregation than at their original level, because of the social processes taking place at the higher level.

Ecological fallacy. Errors committed by taking a relation between variables established at one level and transferring it to a different level without checking its validity for that level.

Intraclass correlation. A measure for the internal homogeneity of level-two units (also called clusters, classes, or groups) with respect to a given variable. It can be interpreted in two equivalent ways: the proportion of total variance of this variable that is accounted for by the higher level; or the correlation between the variables measured for two randomly drawn different individuals in the same group.

Design effect of a two-stage sample. The ratio of the variance of estimation obtained with the given sampling design, to the variance of estimation obtained for a simple random sample from the same population, and with the same total sample size.

Reliability of aggregated variables. Cluster means can be used as measurements for properties of the clusters. Their reliability depends on the number of units used for averaging and on the intraclass correlation.

Within- and between-group regressions. In a multilevel structure, the relation between two variables X and Y will often be different at the different levels of the hierarchy. The within-group regression coefficient of Y on X is the expected difference in the values of Y , when comparing two level-one units in the same group with a unit difference in the value of the variable X . The between-group regression coefficient of Y on X is the expected difference in the values of the group means for Y , when comparing two groups with a unit difference in the group mean with respect to X . The distinction between between-group and within-group regression coefficients is discussed further in Section 4.6. Similarly, within-group correlations are defined as correlations valid within the groups, and the between-group correlations as the correlations between the group means. A population model was introduced to clarify these concepts and show how they relate to total regressions and total correlations,

which are the regressions and correlations obtained for a simple random sample from this population.

Meta-analysis. Statistical methods for the combination of several studies.

Combination of tests. A method for combining results from several studies demanding little in terms of assumptions, only their independence. The best-known method is Fisher's combination of independent p -values.

Combination of estimates. If several studies are combined and it is reasonable to assume that the same parameter is estimated in all of these studies, then a pooled parameter estimate can be obtained with smaller variance than the individual estimates.

4

The Random Intercept Model

In the preceding chapters it was argued that the best approach to the analysis of multilevel data is one that represents within-group as well as between-group relations within a single analysis, where ‘group’ refers to the units at the higher levels of the nesting hierarchy. Very often it makes sense to represent the variability within and between groups by a probability model, in other words, to conceive of the unexplained variation within groups and that between groups as random variability. For a study of students within schools, for example, this means that not only unexplained variation between students, but also unexplained variation between schools is regarded as random variability. This can be expressed by statistical models with so-called random coefficients.

The hierarchical linear model is such a random coefficient model for multilevel, or hierarchically structured, data and has become the main tool for multilevel analysis. In this chapter and the next the definition of this model and the interpretation of the model parameters are discussed. This chapter discusses the simpler case of the random intercept model; Chapter 5 treats the general hierarchical linear model, which also has random slopes. Chapter 6 is concerned with testing the various components of the model. Later chapters treat various elaborations and other aspects of the hierarchical linear model. The focus of this treatment is on the two-level case, but Chapters 4 and 5 also contain sections on models with more than two levels of variability.

OVERVIEW OF THE CHAPTER

This chapter starts by discussing regression models that may have different parameters in different groups, and discusses when it is more appropriate to have separate parameters in each group (‘fixed effects’), and when to have random variability between groups (‘random effects’). Then the random intercept model is defined as a random effects model, with special attention given to its simplest case, the *empty model*, serving as a starting point for multilevel modeling. Here also a measure for heterogeneity between groups is defined, the intraclass correlation coefficient.

A major issue is treated next: the avoidance of ecological fallacies by differentiating the between-group regression from the within-group regression. Then estimation techniques for the parameters are discussed on an intuitive, nonmathematical level. A special topic

is the estimation of the random coefficients in each group by posterior means, where the information for each single group is combined with the information for the whole data set. This is an example of the advantages of the multilevel approach: for inference about any given group, strength is borrowed from the data for other groups. The final section of this chapter is about random intercept models with more than two levels.

4.1 Terminology and notation

For the sake of concreteness, in a two-level model we refer to the level-one units as ‘individuals’, and to the level-two units as ‘groups’. The reader with a different application in mind may supply other names for the units; for example, if the application is to repeated measurements, ‘measurement occasions’ for level-one units and ‘subjects’ for level-two units. The nesting situation of ‘measurement occasions within individuals’ is given special attention in Chapter 15. The groups are also called clusters in the literature. The number of groups in the data is denoted by N ; the number of individuals in the groups may vary from group to group, and is denoted by n_j for group j ($j = 1, 2, \dots, N$). The total number of individuals is denoted by $M = \sum_j n_j$.

The hierarchical linear model is a type of regression model that is particularly suitable for multilevel data. It differs from the usual multiple regression model in that the equation defining the hierarchical linear model contains more than one error term: one (or more) for each level. As in all regression models, a distinction is made between *dependent* and *explanatory* variables: the aim is to construct a model that expresses how the dependent variable depends on, or is explained by, the explanatory variables. Instead of ‘explanatory variable’, the names ‘predictor variable’ and ‘independent variable’ are also used; and ‘criterion’ is also used for ‘dependent variable’. The dependent variable must be a variable at level one: the hierarchical linear model is a model for explaining something that happens at the lowest, most detailed level. Models for multilevel structures where the dependent variable is defined at a higher level are treated by Croon and van Veldhoven (2007), Lüdtke et al. (2008), and van Mierlo et al. (2009).

In this section, we assume that one explanatory variable is available at either level. In the notation, we distinguish the following types of indices and variables:

j is the index for the groups ($j = 1, \dots, N$);

i is the index for the individuals within the groups ($i = 1, \dots, n_j$).

The indices can be regarded as case numbers; note that the numbering for the individuals starts again in every group. For example, individual 1 in group 1 is different from individual 1 in group 2.

For individual i in group j , we have the following variables:

Y_{ij} is the dependent variable

x_{ij} is the explanatory variable at the individual level;

for group j ,

z_j is the explanatory variable at the group level.

To understand the notation, it is essential to realize that the indices i and j indicate precisely what the variables depend on. The notation Y_{ij} , for example, indicates that the value of variable Y depends on group j and also on individual i . (Since the individuals are nested within groups, the index i makes sense only if it is accompanied by the index j : to identify individual $i = 1$, we must know which group we are referring to!) The notation z_j , on the other hand, indicates that the value of Z depends only on group j , and not on individual i .

The basic idea of multilevel modeling is that the outcome variable Y has an individual as well as a group aspect. This also carries through to other level-one variables. The X -variable, although it is a variable at the individual level, may also contain a group aspect. The mean of X in one group may be different from the mean in another group. In other words, X may (and often will) have a positive between-group variance. To put it more generally, the compositions of the various groups with respect to X may differ from one another. It should be kept in mind that explanatory variables that are defined at the individual level often also contain some information about the groups, as discussed in Section 3.1.

4.2 A regression model: fixed effects only

The simplest model is one without the random effects that are characteristic of multilevel models; it is the classical model of multiple regression. This model states that the dependent variable, Y_{ij} , can be written as the sum of a systematic part (a linear combination of the explanatory variables) and a random residual,

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 z_j + R_{ij}. \quad (4.1)$$

In this model equation, the β s are the regression parameters: β_0 is the intercept (i.e., the value obtained if both x_{ij} and z_j are 0), β_1 is the coefficient for the individual variable X , and β_2 is the coefficient for the group variable Z . The variable R_{ij} is the residual (sometimes called *error*) – an essential requirement in regression model (4.1) is that all residuals are mutually independent and have mean 0; a convenient assumption is that in all groups they have the same variances (the homoscedasticity assumption) and are normally distributed. This model has a multilevel nature only to the extent that some explanatory variables may refer to the lower and others to the higher level.

Model (4.1) can be extended to a regression model containing not only the main effects of X and Z , but also the cross-level interaction effect. This type of interaction is discussed in more detail in Chapter 5. It means that the product variable $ZX = Z \times X$ is added to the list of explanatory variables. The resulting regression equation is

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 z_j + \beta_3 z_j x_{ij} + R_{ij}, \quad (4.2)$$

These models pretend, as it were, that all the multilevel structure in the data is fully explained by the group variable Z and the individual variable X . In other words, if two individuals are being considered and their X - and Z -values are given, then for their Y -value it is immaterial whether they belong to the same or to different groups.

Models of types (4.1) and (4.2), and their extensions to more explanatory variables at either or both levels, have in the past been widely used in research on data with a multi-level structure. They are convenient to handle for anybody who knows multiple regression

analysis. Is anything wrong with them? YES! For data with a meaningful multilevel structure, it is practically always incorrect to make the *a priori* assumption that all of the group structure is represented by the explanatory variables. Given that there are only N groups, it is unjustified to act as if one has $n_1 + n_2 + \dots + n_N$ independent replications. There is one exception: when all group sample sizes n_j are equal to 1, the researcher need have no qualms about using these models because the nesting structure is not present in the relation between observed variables, even if it may be present in the structure of the population. Designs with $n_j = 1$ can be used when the explanatory variables were chosen on the basis of substantive theory, and the focus of the research is on the regression coefficients rather than on how the variability of Y is partitioned into within-group and between-group variability.

In designs with group sizes larger than 1, however, the nesting structure often cannot be represented completely in the regression model by the explanatory variables. Additional effects of the nesting structure can be represented by letting the regression coefficients vary from group to group. Thus, the coefficients β_0 and β_1 in equation (4.1) must depend on the group, denoted by j . This is expressed in the formula by an extra index j for these coefficients. This yields the model

$$Y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \beta_2 z_j + R_{ij}. \quad (4.3)$$

Groups j can have a higher (or lower) value of β_{0j} , indicating that, for any given value of X , they tend to have higher (or lower) values of the dependent variable Y . Groups can also have a higher or lower value of β_{1j} , which indicates that the effect of X on Y is higher or lower. Since Z is a group-level variable, it would not make much sense conceptually to let the coefficient of Z depend on the group. Therefore β_2 is left unaltered in this formula.

The multilevel models treated in the following sections and in Chapter 5 contain diverse specifications of the varying coefficients β_{0j} and β_{1j} . The simplest version of model (4.3) is that where β_{0j} and β_{1j} are constant (do not depend on j), that is, the nesting structure has no effect, and we are back at model (4.1). If this is an appropriate model, which we said above is a doubtful supposition, then the ordinary least squares (OLS) regression models of type (4.1) and (4.2) offer a good approach to analyzing the data. If, on the other hand, the coefficients β_{0j} and β_{1j} do depend on j , then these regression models may give misleading results. Then it is preferable to take into account how the nesting structure influences the effects of X and Z on Y . This can be done using the random coefficient model of this and the following chapters. This chapter examines the case where the intercept β_{0j} depends on the group; the next chapter treats the case where the regression coefficient β_{1j} is also group-dependent.

4.3 Variable intercepts: fixed or random parameters?

Let us first consider only the regression on the level-one variable X . A first step toward modeling between-group variability is to let the intercept vary between groups. This reflects the tendency for some groups to have, on average, higher responses Y and others to have lower responses. This model is halfway between (4.1) and (4.3) in the sense that the intercept β_{0j} does depend on the group but the regression coefficient of X , β_1 , is constant:

$$Y_{ij} = \beta_{0j} + \beta_1 x_{ij} + R_{ij}. \quad (4.4)$$

For simplicity here the effect of Z or other variables is omitted, but the discussion applies equally to the case with more explanatory variables.

This model is depicted in Figure 4.1. The group-dependent intercept can be split into an average intercept and the group-dependent deviation:

$$\beta_{0j} = \gamma_{00} + U_{0j}.$$

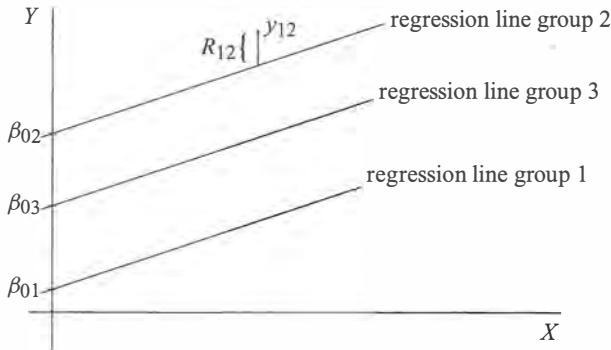


Figure 4.1: Different parallel regression lines. The point y_{12} is indicated with its residual R_{12} .

For reasons that will become clear in Chapter 5, the notation for the regression coefficients is changed here, and the average intercept is called γ_{00} while the regression coefficient for X is called γ_{10} . Substitution now leads to the model

$$Y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + U_{0j} + R_{ij}. \quad (4.5)$$

The values U_{0j} are the main effects of the groups: conditional on an individual having a given X -value and being in group j , the Y -value is expected to be U_{0j} higher than in the average group. Model (4.5) can be understood in two ways:

1. As a model where the U_{0j} are *fixed* parameters, N in number, of the statistical model. This is relevant if the groups j refer to categories each with their own distinct interpretation – for example, a classification according to gender or religious denomination. In order to obtain identified parameters, the restriction that $\sum_j U_{0j} = 0$ can be imposed, so that effectively the groups lead to $N - 1$ regression parameters. This is the usual analysis of covariance model, in which the grouping variable is a factor. (Since we prefer to use Greek letters for statistical parameters and capital letters for random variables, we would prefer to use a Greek letter rather than U when we take this view of model (4.5).) In this specification it is impossible to use a group-level variable Z_j as an explanatory variable because it would be redundant given the fixed group effects.
2. As a model where the U_{0j} are independent and identically distributed *random* variables. Note that the U_{0j} are the unexplained group effects, which also may be called

group residuals, controlling for the effects of variable X . These residuals are now assumed to be randomly drawn from a population with zero mean and an *a priori* unknown variance. This is relevant if the effects of the groups j (which may be neighborhoods, schools, companies, etc.), controlling for the explanatory variables, can be considered to be exchangeable. There is one parameter associated with the U_{0j} in this statistical model: their variance. This is the simplest random coefficient regression model. It is called the *random intercept model* because the group-dependent intercept, $\gamma_{00} + U_{0j}$, is a quantity that varies randomly from group to group. The groups are regarded as a sample from a population of groups. It is possible that there are group-level variables Z_j that express relevant attributes of the groups (such variables will be incorporated in the model in Section 4.5 and, more extensively, in Section 5.2). Then it is not the groups that are exchangeable but the group-level residuals U_{0j} conditional on the group variables.

The next section discusses how to determine which of these specifications of model (4.5) is appropriate in a given situation.

Note that models (4.1) and (4.2) are OLS models, or fixed effects models, which do not take the nesting structure into account (except perhaps by the use of a group-level variable Z_j), whereas models of type (1) above are OLS models that do take the nesting structure into account. The latter kind of OLS model has a much larger number of regression parameters, since in such models N groups lead to $N - 1$ regression coefficients. It is important to distinguish between these two kinds of OLS models in discussing how to handle data with a nested structure.

4.3.1 When to use random coefficient models

These two different interpretations of equation (4.5) imply that multilevel data can be approached in two different ways, using models with fixed or with random coefficients. Which of these two interpretations is the most appropriate in a given situation depends on the focus of the statistical inference, the nature of the set of N groups, the magnitudes of the group sample sizes n_j , and the population distributions involved.

1. If the groups are regarded as unique categories and the researcher wishes primarily to draw conclusions pertaining to each of these N specific categories, then it is appropriate to use the analysis of covariance (fixed effects) model. Examples are groups defined by gender or ethnic background.
2. If the groups are regarded as a sample from a (real or hypothetical) population and the researcher wishes to draw conclusions pertaining to this population, then the random coefficient model is appropriate. Examples are the groupings mentioned in Table 2.2.
3. If the researcher wishes to test effects of group-level variables, the random coefficient model should be used. The reason is that the fixed effects model already ‘explains’ all differences between groups by the fixed effects, and there is no unexplained between-group variability left that could be explained by group-level variables. ‘Random effects’ and ‘unexplained variability’ are two ways of saying the same thing.
4. Especially for relatively small group sizes (in the range from 2 to 50 or 100), the random coefficient model has important advantages over the analysis of covariance

model, provided that the assumptions about the random coefficients are reasonable. This can be understood as follows.

The random coefficient model includes the extra assumption of independent and identically distributed group effects U_{0j} . To put it less formally, the unexplained parts of the group differences are governed by ‘mechanisms’ that are roughly similar from one group to the next, and operate independently between the groups. The group-level residuals are then said to be *exchangeable*. This assumption helps to counteract the paucity of the data that is implied by relatively small group sizes n_j . Since all group effects are assumed to come from the same population, the data from each group also have a bearing on inference with respect to the other groups, namely, through the information that is provided about the population of groups.

In the analysis of covariance model, on the other hand, each of the U_{0j} is estimated as a separate parameter (or fixed effect). If group sizes are small, then the data do not contain very much information about the values of the U_{0j} and there will be considerable overfitting in the analysis of covariance model: many parameters have large standard errors. This overfitting is avoided by using the random coefficient model, because the U_{0j} do not figure as individual parameters. If, on the other hand, the group sizes are large (say, 100 or more), then in the analysis of covariance the group-dependent parameters U_{0j} are estimated very precisely (with small standard errors), and the additional information that they come from the same population does not add much to this precision. In such a situation the difference between the results of the two approaches will be negligible.

5. If the researcher is interested only in within-group differences, and wishes to control as much as possible for between-group differences without modeling them as such or testing them, then the analysis of covariance model is more appropriate. This is because the only assumption made about the group differences is that they can be represented by main group effects β_{0j} as in (4.4), without any assumption about their distribution or about how they might be associated with the explanatory variables in the model. Thus, the parameter β_1 in the fixed effects interpretation of (4.4) represents only the within-group regression parameter: it is the expected difference between values of the dependent variable for cases in the same group differing by one unit on variable X , and having the same values for all other explanatory variables, without any influence of differences between the group averages.
6. The random coefficient model is mostly used with the additional assumption that the random coefficients, U_{0j} and R_{ij} in (4.5), are normally distributed. If this assumption is a very poor approximation, results obtained with this model may be unreliable. This can happen, for example, when there are more outlying groups than can be accounted for by a normally distributed group effect U_{0j} with a common variance. Models for group residuals with nonnormal distributions have been developed, however; see Sections 10.8, 12.2, and 12.3.

Other discussions about the choice between fixed and random coefficients can be found, for example, in Searle et al. (1992; Section 1.4), Hsiao (1995, Section 8), and Greene (2008, Chapter 9). An often mentioned condition for the use of random coefficient models is the restriction that the random coefficients should be independent of the explanatory

variables. However, if there is a possible correlation between group-dependent coefficients and explanatory variables, this residual correlation can be removed, while continuing to use a random coefficient model, by also including effects of the group means of the explanatory variables. Therefore, such a correlation does not imply the necessity of using a fixed effects model instead of a random coefficient model. This is treated in Sections 4.6 and 10.2.1.

In order to choose between regarding the group-dependent intercepts U_{0j} as fixed statistical parameters and regarding them as random variables, a rule of thumb that often works in educational and social research is the following. This rule mainly depends on N , the number of groups in the data. If N is small, say, $N < 10$, then use the analysis of covariance approach: the problem with viewing the groups as a sample from a population in this case is that the data will contain only scant information about this population. If N is not small, say $N \geq 20$; if, furthermore, the level-two units are indeed regarded as a sample of an actual or hypothetical population (see the discussion on analytic inference in Section 14.1.1) and the researcher wishes to draw conclusions about differences in this population (such as regression coefficients for level-two variables or unexplained variability between level-two units); and if also n_j is small or intermediate, say $n_j < 100$ – then use the random coefficient approach. The reason here is that 20 or more groups is usually too large a number to be regarded as unique categories. If the group sizes n_j are large, say all $n_j \geq 100$, then it does not matter much which view we take provided that within- and between-group regression coefficients are appropriately differentiated (Section 4.6). The situation with 10–20 groups is ambiguous in this respect. However, this rule of thumb should be taken with a large pinch of salt and serves only to provide a first hunch, not to determine the choice between fixed and random effects.¹

Populations and populations

Having chosen to work with a random coefficient model, the researcher must be aware that more than one population is involved in the multilevel analysis. Each level corresponds to a population. For example, for a study of students in schools, there is a population of schools and a population of students; for voters in municipalities, there is a population of municipalities and a population of voters. Recall that in this book we take a model-based view (cf. Section 14.1). This implies that populations are infinite hypothetical entities, which express ‘what could be the case’. The random residuals and coefficients can be regarded as representing the effects of unmeasured variables and the approximate nature of the linear model. Randomness, in this sense, may be interpreted as unexplained variability.

Sometimes a random coefficient model can be used also when the population idea at the lower level is less natural. For example, in a study of longitudinal data where respondents are measured repeatedly, a multilevel model can be used with respondents at the second and measurements at the first level: measurements are nested within respondents. Then the population of respondents is an obvious concept. Measurements may be related to a population of time points. This will sometimes be natural, but not always. Another way

¹Maas and Hox (2005) conclude that for $N = 30$ groups even of small sizes $n_j = 5$, the coverage rates of confidence intervals of fixed parameters of the hierarchical linear model are still satisfactory. The fact that for $N = 30$ they obtain unacceptable coverage rates for the asymptotic confidence intervals for variance parameters indicates that, for such a small number of groups, these asymptotic confidence intervals should not be used for variance parameters. This is, however, not an argument against multilevel modeling in general for $N \leq 30$ groups.

of expressing the idea of random coefficient models in such a situation is to say that the level-two units are a sample from a population, residual (unexplained) variability is present at level one as well as at level two, and this unexplained variability is represented as random variation in a probability model.

4.4 Definition of the random intercept model

Here we treat the random coefficient view of model (4.5). This model, the *random intercept model*, is a simple case of the so-called *hierarchical linear model*. We shall not specifically treat the analysis of covariance model, and refer for this purpose to texts on analysis of variance and covariance (Shadish et al., 2002; Stevens, 2009). However, we shall encounter a number of considerations from the analysis of covariance that also play a role in multilevel modeling.

The empty model

Although this chapter follows an approach along the lines of regression analysis, the simplest case of the hierarchical linear model is the *random effects analysis of variance model*, in which the explanatory variables, X and Z , do not figure. This model only contains random groups and random variation within groups. It can be expressed as a model – the same model encountered before in formula (3.1) – where the dependent variable is the sum of a general mean, γ_{00} , a random effect at the group level, U_{0j} , and a random effect at the individual level, R_{ij} :

$$Y_{ij} = \gamma_{00} + U_{0j} + R_{ij}. \quad (4.6)$$

Groups with a high value of U_{0j} tend to have, on average, high responses whereas groups with a low value of U_{0j} tend to have, on average, low responses. The random variables U_{0j} and R_{ij} are assumed to have a mean of 0 (the mean of Y_{ij} is already represented by γ_{00}), to be mutually independent, and to have variances $\text{var}(R_{ij}) = \sigma^2$ and $\text{var}(U_{0j}) = \tau_0^2$. In the context of multilevel modeling (4.6) is called the *empty model*, because it contains not a single explanatory variable. It is important because it provides the basic partition of the variability in the data between the two levels. Given model (4.6), the total variance of Y can be decomposed as the sum of the level-two and level-one variances,

$$\text{var}(Y_{ij}) = \text{var}(U_{0j}) + \text{var}(R_{ij}) = \tau_0^2 + \sigma^2.$$

The covariance between two individuals (i and i' , with $i \neq i'$) in the same group j is equal to the variance of the contribution U_{0j} that is shared by these individuals,

$$\text{cov}(Y_{ij}, Y_{i'j}) = \text{var}(U_{0j}) = \tau_0^2,$$

and their correlation is

$$\rho(Y_{ij}, Y_{i'j}) = \frac{\tau_0^2}{\tau_0^2 + \sigma^2}. \quad (4.7)$$

This parameter is just the *intraclass correlation coefficient* $\rho_I(Y)$ which we encountered in Chapter 3. It can be interpreted in two ways: it is the correlation between two randomly drawn individuals in one randomly drawn group, and it is also the fraction of total variability that is due to the group level.

Example 4.1 Empty model for language scores in elementary schools.

In this example a data set is used that will recur in examples in many subsequent chapters. The data set is concerned with grade 8 students (age about 11 years) in elementary schools in the Netherlands. Further background about the study and primary results can be found in Brandsma and Knuver (1989), Knuver and Brandsma (1993), and Doolaard (1999).

After deleting 258 students with missing values, the number of students is $M = 3,758$, and the number of schools is $N = 211$. Class sizes in the original data set range from 5 to 36. In the data set reduced by deleting cases with missing data, the class sizes range from 4 to 34. The small class sizes are for classes where grade 8 is combined with grade 7 in a larger classroom; here only the grade 8 students are studied. One class per school is included, so the class and the school level are the same in this data set. The nesting structure is students within classes.²

The dependent variable is the score on a language test. Most of the analyses of this data set in this book are concerned with investigating how the language test score depends on the pupil's intelligence and his or her family's socio-economic status, and on related class variables.

Fitting the empty model yields the parameter estimates presented in Table 4.1. The 'deviance' in this table is given for the sake of completeness and later reference; this concept is explained in Chapter 6. In the examples in this and the next chapter, the ML estimation method is used (see Section 4.7) in order to obtain deviances that can be used for comparing all the models by the deviance (likelihood ratio) test of Section 6.2.

Table 4.1: Estimates for empty model.

Fixed effect	Coefficient	S.E.
γ_{00} = Intercept	41.00	0.32
Random part	Variance component	S.E.
<i>Level-two variance:</i>		
$\tau_0^2 = \text{var}(U_{0j})$	18.12	2.16
<i>Level-one variance:</i>		
$\sigma^2 = \text{var}(R_{ij})$	62.85	1.49
Deviance	26,595.3	

The estimates $\hat{\sigma}^2 = 62.85$ and $\hat{\tau}_0^2 = 18.12$ yield an intraclass correlation coefficient of $\hat{\rho}_I = 18.12/80.97 = 0.22$. This is on the high side but not unusual, compared to other results in educational research (values between 0.10 and 0.25 are common – see Hedges and Hedberg, 2007). This indicates

²In the first edition of this book we used a subset of this data set, from which students and schools with missing variables on other variables were also dropped. This was not necessary and may have induced bias; therefore we now use the larger data set. In Chapter 9 the effect of dropping the 258 students with some missing variables is investigated, and turns out to be unimportant. To avoid complications at an early stage, we use here a data set which has no variables with missing data.

that the grouping according to classes leads to an important similarity between the results of different students in the same class, although (as practically always) within-class differences are far larger than between-class differences.

For the overall distribution of the language scores, these estimates provide a mean of 41.00 and a standard deviation of $\sqrt{18.12 + 62.85} = 9.00$. The mean of 41.00 should be interpreted as the expected value of the language score for a random pupil in a randomly drawn class. This is close, but not identical, to the raw mean 41.41 and standard deviation 8.89 of the sample of 3,758 students. The reason for this difference is that the estimation of model (4.6) implies a weighting of the various classes that is not taken into account in the calculation of the raw mean and standard deviation.

The estimates obtained from multilevel software for the two variance components, τ_0^2 and σ^2 , will usually be slightly different from the estimates obtained from the formulas in Section 3.3.1. The reason is that different estimation methods are used: multilevel software uses the more efficient ML or REML method (cf. Section 4.7 below), which in most cases cannot be expressed in an explicit formula; the formulas of Section 3.3.1 are explicit but less efficient.

One explanatory variable

The following step is the inclusion of explanatory variables. These are used to try to explain part of the variability of Y ; this refers to variability at level two as well as level one. With just one explanatory variable X , model (4.5) is obtained (repeated here for convenience):

$$Y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + U_{0j} + R_{ij}.$$

Essential assumptions are that all residuals, U_{0j} and R_{ij} , are mutually independent and have zero means given the values x_{ij} of the explanatory variable. It is assumed that the U_{0j} and R_{ij} are drawn from normally distributed populations. The population variance of the lower-level residuals R_{ij} is assumed to be constant across the groups, and is again denoted by σ^2 ; the population variance of the higher-level residuals U_{0j} is denoted by τ_0^2 . Thus, model (4.5) has four parameters: the regression coefficients γ_{00} and γ_{10} and the variance components σ^2 and τ_0^2 .

The random variables U_{0j} can be regarded as residuals at the group level, or group effects that are left unexplained by X . Since residuals, or random errors, contain those parts of the variability of the dependent variable that are not modeled explicitly as a function of explanatory variables, this model contains unexplained variability at two nested levels. This partition of unexplained variability over the various levels is the essence of hierarchical random effects models.

The fixed intercept γ_{00} is the intercept for the average group. The regression coefficient γ_{10} can be interpreted as an unstandardized regression coefficient in the usual way: a one-unit increase in the value of X is associated with an average increase in Y of γ_{10} units. The residual variance (i.e., the variance conditional on the value of X) is

$$\text{var}(Y_{ij} | x_{ij}) = \text{var}(U_{0j}) + \text{var}(R_{ij}) = \tau_0^2 + \sigma^2,$$

while the covariance between two different individuals (i and i' , with $i \neq i'$) in the same group is

$$\text{cov}(Y_{ij}, Y_{i'j} | x_{ij}, x_{i'j}) = \text{var}(U_{0j}) = \tau_0^2.$$

The fraction of residual variability that can be ascribed to level one is given by $\sigma^2 / (\sigma^2 + \tau_0^2)$, and to level two by $\tau_0^2 / (\sigma^2 + \tau_0^2)$.

Part of the covariance or correlation between two individuals in the same group may be explained by their X -values, and part is unexplained. This unexplained, or residual, correlation between the X -values of these individuals is the *residual intraclass correlation coefficient*,

$$\rho_l(Y|X) = \frac{\tau_0^2}{\sigma^2 + \tau_0^2}.$$

This parameter is the correlation between the Y -values of two randomly drawn individuals in one randomly drawn group, controlling for variable X . It is analogous to the usual intra-class correlation coefficient, but now controlling for X . The formula for the (nonresidual, or raw) intraclass correlation coefficient was just the same, but in that case the variance parameters τ_0^2 and σ^2 referred to the variances in the empty model whereas now they refer to the variances in model (4.5), which includes the effect of variable X . The values of residual intraclass correlations often are smaller than those of raw intraclass correlation coefficients, and in educational research often range between 0.05 and 0.20 (Hedges and Hedberg, 2007).

If model (4.5) is valid while the intraclass correlation coefficient is 0 ($U_{0j} = 0$ for all groups j), then the grouping is irrelevant for the Y -variable conditional on X , and one could have used ordinary linear regression, that is, a model such as (4.1). If the residual intraclass correlation coefficient (or equivalently τ_0^2) is positive, then the hierarchical linear model is a better method of analysis than OLS regression analysis because the standard errors of the estimated coefficients produced by ordinary regression analysis are not to be trusted. This was discussed in Section 3.2.³

The residual intraclass correlation coefficient is positive if and only if the intercept variance τ_0^2 is positive. Therefore testing the residual intraclass correlation coefficient amounts to the same thing as testing the intercept variance. Tests for this parameter are presented in Sections 6.2 and 6.3.

Example 4.2 Random intercept and one explanatory variable: IQ.

As a variable at the pupil level that is essential for explaining language score, we use the measure for verbal IQ taken from the ISI test (Snijders and Welten, 1968). The IQ score has been centered, so that its mean is 0. (The centering took place before taking out the 258 students with some missing values, so that in this data set the mean is close but not exactly equal to 0.) This facilitates interpretation of various parameters. Its standard deviation in this data set is 2.04 (this is calculated as a descriptive statistic, without taking the grouping into account). The results are presented in Table 4.2.

In the model presented in Table 4.2 each class, indexed by the letter j , has its own regression line, given by

$$Y = 41.06 + U_{0j} + 2.507 \text{IQ}.$$

³Mind the word *valid* at the beginning of this paragraph. You may have fitted model (4.5) and obtained an estimated value for $\rho_l(Y)$ that is quite low or even 0. This does not exclude the possibility that there are group differences of the kind treated in Chapter 5: random slopes. If there is an important random slope variance, even without an intercept variance, ordinary regression analysis may yield incorrect results and less insight than the hierarchical linear model.

Table 4.2: Estimates for random intercept model with effect for IQ.

Fixed effect	Coefficient	S.E.
γ_{00} = Intercept	41.06	0.24
γ_{10} = Coefficient of IQ	2.507	0.054
Random part	Variance component	S.E.
<i>Level-two variance:</i> $\tau_0^2 = \text{var}(U_{0j})$	9.85	1.21
<i>Level-one variance:</i> $\sigma^2 = \text{var}(R_{ij})$	40.47	0.96
Deviance	24,912.2	

The U_{0j} are class-dependent deviations of the intercept and have a mean of 0 and a variance of 9.85 (hence, a standard deviation of $\sqrt{9.85} = 3.14$). Figure 4.2 depicts 15 such random regression lines. This figure can be regarded as a random sample from the population of schools defined by Table 4.2.

The scatter around the regression lines (i.e., the vertical distances R_{ij} between the observations and the regression line for the class under consideration) has a variance of 40.47 and therefore a standard deviation of $\sqrt{40.47} = 6.36$. These distances between observations and regression lines therefore tend to be much larger than the vertical distances between the regression lines. However, the distances between the regression lines are not negligible.

A school with a typical low average achievement (bottom 2.5%) will have a value of U_{0j} of about two standard deviations below the expected value of U_{0j} , so that it will have a regression line

$$Y = 41.06 - 2 \times 3.14 + 2.507 \text{ IQ} = 34.78 + 2.507 \text{ IQ},$$

whereas a school with a typical high achievement (top 2.5%) will have a regression line

$$Y = 41.06 + 2 \times 3.14 + 2.507 \text{ IQ} = 47.34 + 2.507 \text{ IQ}.$$

There appears to be a strong effect of IQ. Each additional measurement unit of IQ leads, on average, to 2.507 additional measurement units of the language score. To obtain a scale for effect that is independent of the measurement units, one can calculate *standardized coefficients*, that is, coefficients expressed in standard deviations as scale units. These are the coefficients that would be obtained if all variables were rescaled to unit variances. They are given by

$$\frac{\text{S.D.}(X)}{\text{S.D.}(Y)} \gamma,$$

in this case estimated by $(2.04/9.00) \times 2.507 = 0.57$. In other words, each additional standard deviation on IQ leads, on average, to an increase in language score of 0.57 standard deviations.

The residual variance σ^2 as well as the random intercept variance τ_0^2 are much lower in this model than in the empty model (cf. Table 4.1). The residual variance is lower because between-pupil differences are partially explained. The intercept variance is lower because classes differ in average

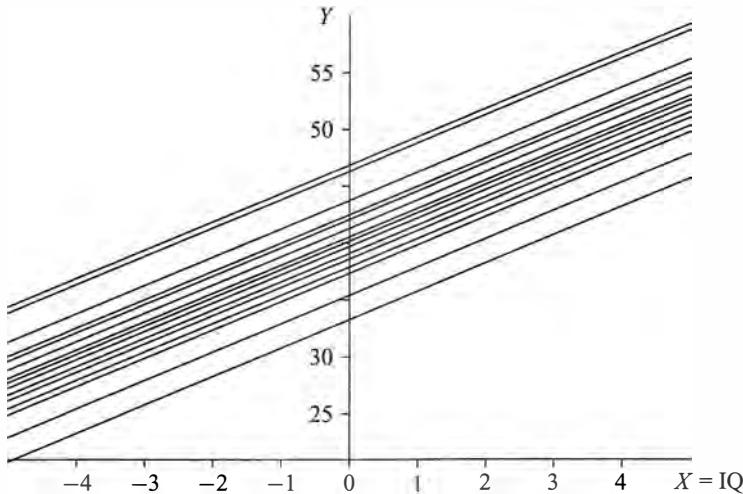


Figure 4.2: Fifteen randomly chosen regression lines according to the random intercept model of Table 4.2.

IQ score, so that this pupil-level variable also explains part of the differences between classes. The residual intraclass correlation is estimated by

$$\hat{\rho}_I(Y|X) = \frac{9.85}{40.47 + 9.85} = 0.20,$$

slightly smaller than the raw intraclass correlation of 0.22 (see Table 4.1).

These results may be compared to those obtained from an OLS regression analysis, in which the nesting of students in classes is not taken into account. This analysis can be regarded as an analysis using model (4.5) in which the intercept variance τ_0^2 is constrained to be 0. The results are displayed in Table 4.3. The parameter estimates for the OLS method seem rather close to those for the random intercept model. However, the regression coefficient for IQ differs by about three standard errors between the two models. This implies that, although the numerical values seem similar, they are nevertheless rather different from a statistical point of view. Further, the standard error of the intercept is twice as large in the results from the random intercept model as in those from the OLS analysis. This indicates that the OLS analysis produces an over-optimistic impression of the precision of this estimate, and illustrates the lack of trustworthiness of OLS estimates for multilevel data.

4.5 More explanatory variables

Just as in multiple regression analysis, more than one explanatory variable can be used in the random intercept model. When the explanatory variables at the individual level are denoted by X_1, \dots, X_p , and those at the group level Z_1, \dots, Z_q , adding their effects to the random intercept model leads to the formula⁴

$$Y_{ij} = \gamma_{00} + \gamma_{10}x_{1ij} + \dots + \gamma_{p0}x_{pj} + \gamma_{01}z_{1j} + \dots + \gamma_{0q}z_{qj} + U_{0j} + R_{ij}. \quad (4.8)$$

⁴The subscripts on the regression coefficients γ may seem somewhat baroque. The reason is to obtain consistency with the notation in the next chapter; see (5.12).

Table 4.3: Estimates for OLS regression.

Fixed effect	Coefficient	S.E.
γ_{00} = Intercept	41.30	0.12
γ_{10} = Coefficient of IQ	2.651	0.056
Random part	Variance component	S.E.
<i>Level-one variance:</i>		
$\sigma^2 = \text{var}(R_{ij})$	49.80	1.15
Deviance	25,351.0	

The regression parameters, γ_{h0} ($h = 1, \dots, p$) and γ_{0h} ($h = 1, \dots, q$) for level-one and level-two explanatory variables, respectively, again have the same interpretation as unstandardized regression coefficients in multiple regression models: a one-unit increase in the value of X_h (or Z_h) is associated with an average increase in Y of γ_{h0} (or γ_{0h}) units. Just as in multiple regression analysis, some of the variables X_h and Z_h may be interaction variables, or nonlinear (e.g., quadratic) transforms of basic variables.

The first part of the right-hand side of (4.8), incorporating the regression coefficients,

$$\gamma_{00} + \gamma_{10}x_{1ij} + \dots + \gamma_{p0}x_{pij} + \gamma_{01}z_{1j} + \dots + \gamma_{0q}z_{qj},$$

is called the *fixed part* of the model, because the coefficients are fixed (i.e., nonstochastic). The remainder,

$$U_{0j} + R_{ij},$$

is called the *random part* of the model.

It is again assumed that all residuals, U_{0j} and R_{ij} , are mutually independent and have zero means given the values of the explanatory variables. A somewhat less crucial assumption is that these residuals are drawn from normally distributed populations. The population variance of the level-one residuals R_{ij} is denoted by σ^2 , while the population variance of the level-two residuals U_{0j} is denoted by τ_0^2 .

Due to the two-level nature of the model under consideration, two special kinds of variables deserve special attention: group-level variables that arise as aggregates of individual-level variables; and cross-level interactions. The latter will be treated in Chapter 5; the former are treated here and in the next section.

Some level-two variables are directly defined for the units of this level; others are defined through the subunits (units at level one) that are comprised in this unit. For example, when the nesting structure refers to children within families, the type of dwelling is directly a family-dependent variable, but the average age of the children is based on aggregation of a variable (age) that is itself a level-one variable. As another example, referring to a longitudinal study of primary school children, where the nesting structure refers to repeated measurements within individual children, the gender of the child is a direct level-two variable, whereas the average reading score over the age period of 6–12 years, or the slope of

the increase in reading score over this period, is a level-two variable that is an aggregate of a level-one variable.

Varying group sizes

In most research, the group sizes n_j are variable between groups. Even in situations such as repeated measures designs, where level two corresponds to individual subjects and level one to time points of measurement of these subjects, and where it is often the researcher's intention to measure all subjects at the same moments, there often occur missing data which again leads to a data structure with variable n_j . This in no way constitutes a problem for the application of the hierarchical linear model. The hierarchical linear model can even be applied if some groups have size $n_j = 1$, as long as some other groups have greater sizes. Of course, the smaller groups will have a smaller influence on the results than the larger groups.

However, it may be worthwhile to give some thought to the substantive meaning of the group size. The number of students in a school class may have effects on teaching and learning processes, the size of a department in a firm may have influence on the performance of the department and on interpersonal relations. In such cases it may be advisable to include group size in the model as an explanatory variable with a fixed effect.

4.6 Within- and between-group regressions

Group means are an especially important type of explanatory variable. A group mean for a given level-one explanatory variable is defined as the mean over all individuals, or level-one units, within the given group, or level-two unit.⁵ This can be an important contextual variable. The group mean of a level-one explanatory variable allows the difference between *within-group* and *between-group* regressions to be expressed, as proposed by Davis et al. (1961) and extended by various authors, including Neuhaus and Kalbfleisch (1998). We saw in Section 3.6 that the coefficients for these two types of regression can be completely different. The within-group regression coefficient expresses the effect of the explanatory variable within a given group; the between-group regression coefficient expresses the effect of the group mean of the explanatory variable on the group mean of the dependent variable. In other words, the between-group regression coefficient is just the coefficient in a regression analysis for data that are aggregated (by averaging) to the group level.

Continuing with the example of children within families, suppose that we are interested in the amount of pocket money that children receive. This will depend on the child's age, but it could also depend on the average age of the children in the family. The within-group regression coefficient measures the effect of age differences within a given family; the between-group regression coefficient measures the effect of average age on the average pocket money received by the children in the family. In a simple model, not taking other

⁵If cases with missing data are deleted in order to carry out the multilevel analysis, then one should calculate the group mean as the average over all level-one units for which this particular variable is available, and before deleting cases because they have missing values on other variables. But instead of deleting cases it is better to use the methods of Chapter 9.

variables into account, denote age (in years) of child i in family j by x_{ij} , and the average age of all children in family j by $z_j = \bar{x}_j$. In the model

$$Y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + U_{0j} + R_{ij},$$

the within-group and between-group regression coefficients are forced to be equal. If we add the family mean, $z_j = \bar{x}_j$, as an explanatory variable, we obtain

$$Y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + \gamma_{01} \bar{x}_j + U_{0j} + R_{ij}. \quad (4.9)$$

This model is more flexible in that the within-group regression coefficient is allowed to differ from the between-group regression coefficient. This can be seen as follows.

If model (4.9) is considered within a given group, the terms can be reordered as

$$Y_{ij} = (\gamma_{00} + \gamma_{01} \bar{x}_j + U_{0j}) + \gamma_{10} x_{ij} + R_{ij}.$$

The part in parentheses is the (random) intercept for this group, and the regression coefficient of X within this group is γ_{10} . The systematic (nonrandom) part for group j is the within-group regression line

$$Y = (\gamma_{00} + \gamma_{01} \bar{x}_j) + \gamma_{10} x.$$

On the other hand, taking the group average on both sides of the equality sign in (4.9) yields the between-group regression model,

$$\begin{aligned} \bar{Y}_j &= \gamma_{00} + \gamma_{10} \bar{x}_j + \gamma_{01} \bar{x}_j + U_{0j} + \bar{R}_j \\ &= \gamma_{00} + (\gamma_{10} + \gamma_{01}) \bar{x}_j + U_{0j} + \bar{R}_j. \end{aligned}$$

The systematic part of this model is represented by the between-group regression line

$$Y = \gamma_{00} + (\gamma_{10} + \gamma_{01}) x.$$

This shows that the between-group regression coefficient is $\gamma_{10} + \gamma_{01}$.

The difference between the within-group and the between-group regression coefficients can be tested in this model by testing the null hypothesis that $\gamma_{01} = 0$ by the method of Section 6.1.

This test is a version of what is known in econometrics as the Hausman specification test; see Hausman and Taylor (1981) or Baltagi (2008, Section 4.3). Some textbooks suggest the use of this test as the primary means of choosing between a fixed effects and a random effects approach; see, for example, Greene (2008, Section 9.5.4). However, if this test is significant, it is not necessary to conclude that one should abandon the use of a random effects model; there may be reasons to continue using a random effects model while including the group mean of X as an explanatory variable. The choice between fixed and random effects models should be based on other considerations, as reviewed in Section 4.3. This incorrect interpretation of the Hausman test is discussed in Fielding (2004a) and Snijders and Berkhof (2008, p. 147).

Figure 4.3 is a sketch of within-group regression lines that differ from the between-group regression line in the sense that the between-group regression is stronger ($\gamma_{01} > 0$). The reverse, where the between-group regression line is less steep than the within-group regression lines, can also occur.

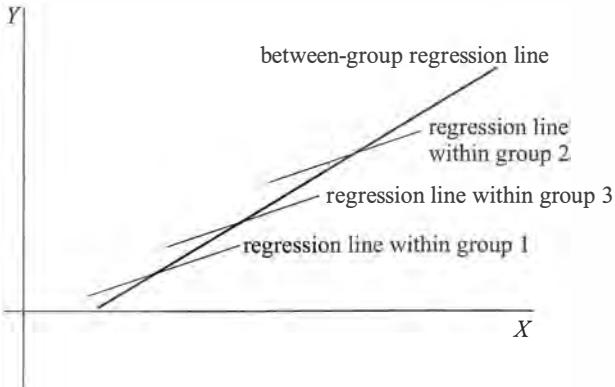


Figure 4.3: Different between-group and within-group regression lines.

If the within- and between-group regression coefficients are different, then it is often convenient to replace x_{ij} in (4.9) with the *within-group deviation score*, defined as $x_{ij} - \bar{x}_j$. To distinguish the corresponding parameters from those in (4.9), they are denoted by $\tilde{\gamma}$. The resulting model is

$$Y_{ij} = \tilde{\gamma}_{00} + \tilde{\gamma}_{10}(x_{ij} - \bar{x}_j) + \tilde{\gamma}_{01}\bar{x}_j + U_{0j} + R_{ij}, \quad (4.10)$$

This model is statistically equivalent to model (4.9) but has a more convenient parametrization because the between-group regression coefficient is now

$$\tilde{\gamma}_{01} = \gamma_{10} + \gamma_{01}, \quad (4.11)$$

while the within-group regression coefficient is

$$\tilde{\gamma}_{10} = \gamma_{10}. \quad (4.12)$$

The use of the within-group deviation score is called *within-group centering*; some computer programs for multilevel analysis have special facilities for this.

Example 4.3 Within- and between-group regressions for IQ.

We continue Example 4.2 by allowing differences between the within-group and between-group regressions of the language score on IQ. The results are displayed in Table 4.4. IQ here is the variable with overall centering but no group centering. Thus, the results refer to model (4.9).

The within-group regression coefficient is 2.454 and the between-group regression coefficient is $2.454 + 1.312 = 3.766$. A pupil with a given IQ obtains, on average, a higher language test score if he or she is in a class with a higher average IQ. In other words, the contextual effect of mean IQ in the class gives an additional contribution over and above the effect of individual IQ. The effect of classroom average IQ is about half that of individual IQ. A graph of these results would be qualitatively similar to Figure 4.3 in the sense that the within-group regression lines are less steep than the between-group regression line.

Table 4.4 represents within each class, denoted j , a linear regression equation

$$Y = 41.11 + U_{0j} + 2.454 \text{ IQ} + 1.312 \bar{I}Q,$$

Table 4.4: Estimates for random intercept model with different within- and between-group regressions.

Fixed effect	Coefficient	S.E.
γ_{00} = Intercept	41.11	0.23
γ_{10} = Coefficient of IQ	2.454	0.055
γ_{01} = Coefficient of \bar{IQ} (group mean)	1.312	0.262
Random part	Variance component	S.E.
<i>Level-two variance:</i>		
$\tau_0^2 = \text{var}(U_{0j})$	8.68	1.10
<i>Level-one variance:</i>		
$\sigma^2 = \text{var}(R_{ij})$	40.43	0.96
Deviance	24,888.0	

where U_{0j} is a class-dependent deviation with mean 0 and variance 8.68 (standard deviation 2.95). The within-class deviations about this regression equation, R_{ij} , have a variance of 40.43 (standard deviation 6.36). Within each class, the effect (regression coefficient) of IQ is 2.454, so the regression lines are parallel. Classes differ in two ways: they may have different mean IQ values, which affects the expected results Y through the term $1.312 \bar{IQ}$; this is an explained difference between the classes; and they have randomly differing values for U_{0j} , which is an unexplained difference. These two ingredients contribute to the class-dependent intercept, given by $41.11 + U_{0j} + 1.312 \bar{IQ}$.

The within-group and between-group regression coefficients would be equal if, in formula (4.9), the coefficient of average IQ were 0 (i.e., $\gamma_{01} = 0$). This null hypothesis can be tested (see Section 6.1) by the t -ratio defined as

$$t = \frac{\text{estimate}}{\text{standard error}},$$

given here by $1.312/0.262 = 5.01$, a highly significant result. In other words, we may conclude that the within- and between-group regression coefficients are indeed different.

If the individual IQ variable had been replaced by within-group deviation scores $IQ_{ij} - \bar{IQ}_j$, that is, model (4.10) had been used, then the estimates obtained would have been $\hat{\gamma}_{10} = 2.454$ and $\hat{\gamma}_{01} = 3.766$; cf. formulas (4.11) and (4.12). Indeed, the regression equation given above can be described equivalently by

$$Y = 41.11 + U_{0j} + 2.454(IQ - \bar{IQ}) + 3.766\bar{IQ},$$

which indicates explicitly that the within-group regression coefficient is 2.454, while the between-group regression coefficient (i.e., the coefficient of the group means \bar{Y} on the group means \bar{IQ}) is 3.766.

When interpreting the results of a multilevel analysis, it is important to keep in mind that the conceptual interpretation of within-group and between-group regression coefficients usually is completely different. These two coefficients may express quite contradictory mechanisms. This is related to the shift of meaning and the ecological fallacy discussed

in Section 3.1. For theoretically important variables in multilevel studies, it is the rule rather than the exception that within-group regression coefficients differ from between-group regression coefficients (although the statistical significance of this difference may be another matter, depending as it does on sample sizes, etc.).

4.7 Parameter estimation

The random intercept model (4.8) is defined by its statistical parameters: the regression parameters, γ , and the variance components, σ^2 and τ_0^2 . Note that the random effects, U_{0j} , are not parameters in a statistical sense, but latent (i.e., not directly observable) variables. The literature (e.g., Longford, 1993; de Leeuw and Meijer, 2008b) contains two major estimation methods for estimating the statistical parameters, under the assumption that the U_{0j} as well as the R_{ij} are normally distributed: maximum likelihood (ML) and residual (or restricted) maximum likelihood (REML).

The two methods differ little with respect to estimating the regression coefficients, but they do differ with respect to estimating the variance components. A very brief indication of the difference between the two methods is that REML estimates the variance components while taking into account the loss of degrees of freedom resulting from the estimation of the regression parameters, while ML does not take this into account. The result is that the ML estimators for the variance components have a downward bias, and the REML estimators do not. For example, the usual variance estimator for a single-level sample, in which the sum of squared deviations is divided by the sample size minus 1, is a REML estimator; dividing instead by the total sample size gives the corresponding ML estimator.

The relative difference between ML and REML estimates will usually be largest for the estimated variance parameters (and covariance parameters for random slope models as treated in Chapter 5). In a first approximation, for a two-level random intercept model the ratio of ML to REML estimates for the random intercept variance may be expected to be something like $N/(N - q - 1)$, where q is the number of level-two explanatory variables (to which should be added a fractional part of level-one explanatory variables that have a strong intraclass correlation). The difference will be important when $N - q - 1$ is small. For a large number (as a rule of thumb, ‘large’ here means $N - q - 1 \geq 50$), the difference between the ML and the REML estimates will be immaterial. The literature (e.g., McCulloch and Searle, 2001, Section 6.10) suggests that the REML method is preferable with respect to the estimation of the variance parameters (and the covariance parameters for the more general models treated in Chapter 5). When one wishes to carry out deviance tests (see Section 6.2), however, the use of ML rather than REML estimates is sometimes required.⁶ For this reason, the estimates for the examples in this chapter and the next have been calculated by the ML method. On the other hand, we shall see in Section 6.1 that for tests of fixed effects it is better, for small values of N , to use standard errors produced by the REML method.

⁶When models are compared with different fixed parts, deviance tests should be based on ML estimation. Deviance tests with REML estimates may be used for comparing models with different random parts and the same fixed part. Different random parts will be treated in the next chapter.

Example 4.4 *ML and REML estimates.*

Example 4.3 presents ML estimates. The REML estimates differ from these in the second or third decimal place, with the exception of the random intercept variance. The ML estimate is 8.68 (standard error 1.10), the REML estimate 8.78 (standard error 1.11). There are 211 classes in the data set and there is $q = 1$ level-one explanatory variable, average IQ. This leads to the expectation that the REML estimate will be roughly $209/211 = 0.99$ times the ML estimate. Indeed, $8.68/8.78 = 0.99$ (the difference is in the third decimal place). The difference of $8.78 - 8.68 = 0.10$ between the estimates is less than one tenth of the standard error, and therefore negligible in practice.

Various algorithms are available to determine these estimates. They have names such as expectation–maximization (EM), Fisher scoring, iterative generalized least squares (IGLS), and residual or restricted IGLS (RIGLS). They are iterative, which means that a number of steps are taken in which a provisional estimate comes closer and closer to the final estimate. When all goes well, the steps converge to the ML or REML estimate. Technical details can be found, for example, in Raudenbush and Bryk (2002), Goldstein (2011), Longford (1993, 1995), and de Leeuw and Meijer (2008b). In principle, the algorithms all yield the same estimates for a given estimation method (ML or REML). The differences are that for some complicated models, the algorithms may vary in the extent of computational problems (sometimes one algorithm may converge and the other not), and the amount of computing time required. For the practical user, the differences between the algorithms are hardly worth thinking about.

An aspect of the estimation of hierarchical linear model parameters that surprises some users of this model is the fact that it is possible that the variance parameters, in model (4.8) notably parameter τ_0^2 , can be estimated to be exactly 0. The value of 0 is then also reported for the standard error by many computer programs! Even when the estimate is $\hat{\tau}_0^2 = 0$, this does not mean that the data imply absolute certainty that the population value of τ_0^2 is equal to 0. Such an estimate can be understood as follows. For simplicity, consider the empty model, (4.6). The level-one residual variance σ^2 is estimated by the pooled within-group variance. The parameter τ_0^2 is estimated by comparing this within-group variability to the between-group variability. The latter is determined not only by τ_0^2 but also by σ^2 , since

$$\text{var}(\bar{Y}_j) = \tau_0^2 + \frac{\sigma^2}{n_j}. \quad (4.13)$$

Note that τ_0^2 , being a variance, cannot be negative. This implies that, even if $\tau_0^2 = 0$, a positive between-group variability is expected. If observed between-group variability is equal to or smaller than what is expected from (4.13) for $\tau_0^2 = 0$, then the estimate $\hat{\tau}_0^2 = 0$ is reported (cf. the discussion following (3.11)).

If the group sizes n_j are variable, the larger groups will, naturally, have a larger influence on the estimates than the smaller groups. The influence of group size on the estimates is, however, mediated by the intraclass correlation coefficient. Consider, for example, the estimation of the mean intercept, γ_{00} . If the residual intraclass correlation is 0, the groups have an influence on the estimated value of γ_{00} that is proportional to their size. In the extreme case where the residual intraclass correlation is 1, each group has an equally large influence, independent of its size. In practice, where the residual intraclass correlation is between 0 and 1, larger groups will have a larger influence, but less than proportionately.

4.8 ‘Estimating’ random group effects: posterior means

The random group effects U_{0j} are latent variables rather than statistical parameters, and therefore are not estimated as an integral part of the statistical parameter estimation. However, there may be many reasons why it is nonetheless desirable to ‘estimate’ them.⁷ This can be done by a method known as *empirical Bayes estimation* which produces so-called *posterior means*; see, for example, Efron and Morris (1975), Gelman et al. (2004, Section 5.4). The basic idea of this method is that U_{0j} is ‘estimated’ by combining two kinds of information: the data from group j ; and the fact (or, rather, the model assumption) that the unobserved U_{0j} is a random variable just like all other random group effects, and therefore has a normal distribution with mean 0 and variance τ_0^2 . In other words, data information is combined with population information.

The formula is given here only for the empty model, that is, the model without explanatory variables. The idea for more complicated models is analogous; formulas can be found in the literature, for example, Longford (1993, Section 2.10) and Snijders and Berkhof (2008, Section 3.3.3).

The empty model was formulated in (4.6) as

$$Y_{ij} = \beta_{0j} + R_{ij} = \gamma_{00} + U_{0j} + R_{ij}.$$

Since γ_{00} is already an estimated parameter, an estimate for β_{0j} will be the same as an estimate for $U_{0j} + \gamma_{00}$. Therefore, estimating β_{0j} and estimating U_{0j} are equivalent problems given that an estimate for γ_{00} is available.

If we only used group j , β_{0j} would be estimated by the group mean, which is also the OLS estimate,

$$\hat{\beta}_{0j} = \bar{Y}_j. \quad (4.14)$$

If we looked only at the population, we would estimate β_{0j} by its population mean, γ_{00} . This parameter is estimated by the overall mean,

$$\hat{\gamma}_{00} = \bar{Y}_\cdot = \sum_{j=1}^N \frac{n_j}{M} \bar{Y}_j,$$

where $M = \sum_j n_j$ denotes the total sample size. Another possibility is to combine the information from group j with the population information. The optimal combined ‘estimate’ for β_{0j} is a weighted average of the two previous estimates,

$$\hat{\beta}_{0j}^{\text{EB}} = \lambda_j \hat{\beta}_{0j} + (1 - \lambda_j) \hat{\gamma}_{00}, \quad (4.15)$$

where EB stands for ‘empirical Bayes’ and the weight λ_j is defined as the reliability of the mean of group j (see (3.21)),

$$\lambda_j = \frac{\tau_0^2}{\tau_0^2 + \sigma^2/n_j}.$$

⁷The word ‘estimate’ is in quotation marks because the proper statistical term for finding likely values of the U_{0j} , being random variables, is *prediction*. The term ‘estimation’ is reserved for finding likely values for statistical parameters. Since prediction is associated in everyday speech, however, with determining something about the future, we prefer to speak here about ‘estimation’ between parentheses.

The ratio of the two weights, $\lambda_j/(1 - \lambda_j)$, is just the ratio of the true variance τ_0^2 to the error variance σ^2/n_j . In practice we do not know the true values of the parameters σ^2 and τ_0^2 , and we substitute estimated values to calculate (4.15).

Formula (4.15) is called the posterior mean or empirical Bayes estimate, for β_{0j} . This term comes from Bayesian statistics. It refers to the distinction between the *prior* knowledge about the group effects, which is based only on the population from which they are drawn, and the *posterior* knowledge which is based also on the observations made about this group. There is an important parallel between random coefficient models and Bayesian statistical models, because the random coefficients used in the hierarchical linear model are analogous to the random parameters that are essential in the Bayesian statistical paradigm. See Gelman et al. (2004, Chapter 1 and Section 5.4).

Formula (4.15) can be regarded as follows. The OLS estimate (4.14) for group j is pushed slightly toward the general mean $\hat{\gamma}_{00}$. This is an example of *shrinkage to the mean* just as is used (e.g., in psychometrics) for the estimation of true scores. The corresponding estimator is sometimes called the Kelley estimator; see, for example, Kelley (1927), Lord and Novick (1968), or other textbooks on classical psychological test theory. From the definition of the weight λ_j it is apparent that the influence of the data of group j itself becomes larger as group size n_j becomes larger. For large groups, the posterior mean is practically equal to the OLS estimate $\hat{\beta}_{0j}$, the intercept that would be estimated from data on group j alone.

In principle, the OLS estimate (4.14) and the empirical Bayes estimate (4.15) are both sensible procedures for estimating the mean of group j . The former is an unbiased⁸ estimate, and does not require the assumption that group j is a random element from the population of groups. The latter is biased toward the population mean, but for a randomly drawn group it has a smaller mean squared error. The squared error averaged over all groups will be smaller for the empirical Bayes estimate, but the price is a conservative (drawn to the average) appraisal of the groups with truly very high or very low values of β_{0j} . The estimation variance of the empirical Bayes estimate is

$$\text{var}(\hat{\beta}_{0j}^{\text{EB}} - \beta_{0j}) = (1 - \lambda_j) \tau_0^2, \quad (4.16)$$

if the uncertainty due to the estimation of γ_{00} and λ_j is neglected, which is a good approximation for medium and large level-two sample sizes N . This formula also is well known from classical psychological test theory (e.g., Lord and Novick, 1968).

The same principle can be applied (but with more complicated formulas) to the ‘estimation’ of the group-dependent intercept $\beta_{0j} = \gamma_{00} + U_{0j}$ in random intercept models that do include explanatory variables, such as (4.8). This intercept can be ‘estimated’ again by γ_{00} plus the posterior mean of U_{0j} , and is then also referred to as the *posterior intercept*.

Instead of being primarily interested in the intercept as defined by $\gamma_{00} + U_{0j}$, which is the value of the regression equation for *all* explanatory variables having value 0, one may also be interested in the value of the regression line for group j for the case where only the level-one variables x_{1ij}, \dots, x_{pij} are 0 while the level-two variables have the values proper to this group. To ‘estimate’ this version of the intercept of group j , we use

$$\hat{\gamma}_{00} + \hat{\gamma}_{01} z_{1j} + \dots + \hat{\gamma}_{0q} z_{qj} + \hat{U}_{0j}^{\text{EB}}, \quad (4.17)$$

⁸Unbiasedness means that the average of many – hypothetical – independent replications of this estimate for this particular group j would be very close to the true value β_{0j} .

where the values $\hat{\gamma}$ indicate the (ML or REML) estimates of the regression coefficients. The values (4.17) also are sometimes called posterior intercepts.

The posterior means (4.15) can be used, for example, to see which groups have unexpectedly high or low values on the outcome variable, given their values on the explanatory variables. They can also be used in a residual analysis, for checking the assumption of normality for the random group effects, and for detecting outliers (see Chapter 10). The posterior intercepts (4.17) indicate the total main effect of group j , controlling for the level-one variables X_1, \dots, X_p , but including the effects of the level-two variables Z_1, \dots, Z_q . For example, in a study of students in schools where the dependent variable is a relevant indicator of scholastic performance, these posterior intercepts could be valuable information for the parents indicating the contribution of the various schools to the performance of their beloved children.

Example 4.5 Posterior means for random data.

We can illustrate the ‘estimation’ procedure by returning to the random digits table (Chapter 3, Table 3.1). Macro-unit 04 in that table has an average of $\bar{Y}_j = 31.5$ over its 10 random digits. The grand mean of the total 100 random digits is $\bar{Y}_{..} = 47.2$. The average of macro-unit 04 thus seems to be far below the grand mean. But the reliability of this mean is only $\lambda_j = 26.7/\{26.7+(789.7/10)\} = 0.25$. Applying (4.15), the posterior mean is calculated as

$$0.25 \times 31.5 + (1 - 0.25) \times 47.2 = 43.3.$$

In words, the posterior mean for macro-unit 04 is 75% (i.e., $1 - \lambda_j$) determined by the grand mean of 47.2 and only 25% (i.e., λ_j) by its OLS mean of 31.5. The shrinkage to the grand mean is evident. Because of the low estimated intraclass correlation of $\hat{\rho}_I = 0.03$ and the low number of observations per macro-unit, $n_j = 10$, the empirical Bayes estimate of the average of macro-unit 04 is closer to the grand mean than to the group mean. In this case this is a clear improvement: there is no between-group variance in the population, and the posterior mean is much closer to the true value of $\gamma_{00} + U_{0j} = 49.5 + 0 = 49.5$ than the group average.

4.8.1 Posterior confidence intervals

Now suppose that parents have to choose a school for their children, and that they wish to do so on the basis of the value a school adds to abilities that students already have when entering the school (as indicated by an IQ test). Let us focus on the language scores. ‘Good’ schools are schools where students on average are ‘over-achievers’, that is to say, they achieve more than expected on the basis of their IQ. ‘Poor’ schools are schools where students on average have language scores that are lower than one would expect given their IQ scores.

In this case the level-two residuals U_{0j} from a two-level model with language as the dependent and IQ as the predictor variable convey the relevant information. But remember that each U_{0j} has to be estimated from the data, and that there is sampling error associated with each residual, since we work with a sample of students from each school. Of course we might argue that within each school the entire population of students is studied, but in general we should handle each parameter estimate with its associated uncertainty since we are now considering the performance of the school for a hypothetical new pupil at this school.

Therefore, instead of simply comparing schools on the basis of the level-two residuals it is better to compare these residuals taking account of the associated confidence intervals.

These must be constructed using the standard errors of the empirical Bayes estimates. Here we must distinguish between two kinds of standard error. The *comparative standard error* expresses the deviation from the unobserved random variable U_{0j} , and is the standard deviation of

$$\hat{U}_{0j}^{\text{EB}} - U_{0j}.$$

The *diagnostic standard error* expresses the deviation from the value 0, which is the overall mean of the variables U_{0j} , and is the standard deviation of

$$\hat{U}_{0j}^{\text{EB}} - 0 = \hat{U}_{0j}^{\text{EB}}.$$

The comparative standard error, also called the *posterior standard deviation* of U_{0j} , is used to assess how well the unobserved level-two contributions U_{0j} can be ‘estimated’ from the data, for example, to compare groups (more generally, level-two units) with each other, as is further discussed later in this section. The diagnostic standard error is used when the posterior means must be standardized for model checking. Using division by the diagnostic standard errors (and if we may ignore the fact that these are themselves subject to error due to estimation), the standardized empirical Bayes estimators can be regarded as standardized residuals having a standard normal distribution if the model is correct, and this can be used for model checking. This is discussed in Chapter 10.

A theoretically interesting property is that the sum of the diagnostic and comparative variances is equal to the random coefficient variance:

$$\text{comparative variance} + \text{diagnostic variance} = \text{var}(\hat{U}_{0j}^{\text{EB}} - U_{0j}) + \text{var}(\hat{U}_{0j}^{\text{EB}}) = \tau_0^2. \quad (4.18)$$

This is shown more generally for hierarchical linear models by Snijders and Berkhof (2008, Section 3.3.3). The interpretation is that, to the extent that we are better able to estimate the random coefficients (smaller comparative standard errors) – for example, because of larger group sizes n_j – the variability of the estimated values \hat{U}_{0j}^{EB} will increase (larger diagnostic standard errors) because they capture more of the true variation between the coefficients U_{0j} .

The comparative standard error of the empirical Bayes estimate is smaller than the root mean squared error of the OLS estimate based on the data only for the given macro-unit (the given school, in our example). This is just the point of using the empirical Bayes estimate. For the empty model the comparative standard error is the square root of (4.16), which can also be expressed as

$$\text{C.S.E.}(\hat{\beta}_{0j}^{\text{EB}}) = \frac{1}{\sqrt{\tau_0^{-2} + n_j \sigma^{-2}}}. \quad (4.19)$$

This formula was also given by Longford (1993, Section 1.7). Thus, the standard error depends on the within-group as well as the between-group variance and on the number of sampled students for the school. For models with explanatory variables, the standard error can be obtained from computer output of multilevel software. Denoting the standard error

for school j shortly by S.E. $_j$, the corresponding 90% confidence intervals can be calculated as the intervals

$$(\hat{\beta}_{0j}^{\text{EB}} - 1.64 \times \text{S.E.}_j, \hat{\beta}_{0j}^{\text{EB}} + 1.64 \times \text{S.E.}_j).$$

Two cautionary remarks are in order, however.

First, the shrinkage construction of the empirical Bayes estimates implies a bias: ‘good’ schools (with a high U_{0j}) will tend to be represented too negatively, ‘poor’ schools (with a low U_{0j}) will tend to be represented too positively (especially if the sample sizes are small). The smaller standard error is bought at the expense of this bias. These confidence intervals have the property that, on average, the random group effects U_{0j} will be included in the confidence interval for 90% of the groups. But for close to average groups the coverage probability is higher, while for the groups with very low or very high group effects the coverage probability will be lower than 90%.

Second, users of such information generally wish to compare a series of groups. This problem was addressed by Goldstein and Healy (1995). The parent in our example will make her or his own selection of schools, and (if the parent is a trained statistician) will compare the schools on the basis of whether the confidence intervals overlap. In that case, the parent is implicitly performing a series of statistical tests on the differences between the group effects U_{0j} . Goldstein and Healy (1995, p. 175) write: ‘It is a common statistical misconception to suppose that two quantities whose 95% confidence intervals just fail to overlap are significantly different at the 5% significance level’. The reader is referred to their article on how to adjust the width of the confidence intervals in order to perform such significance testing. For example, testing the equality of a series of level-two residuals at the 5% significance level requires confidence intervals that are constructed by multiplying the standard error given above by 1.39 rather than the well-known 5% value of 1.96. For a 10% significance level, the factor is 1.24 rather than 1.64. So the ‘comparative confidence intervals’ are allowed to be narrower than the confidence intervals used for assessing single groups.

Example 4.6 Comparing added value of schools.

Table 4.4.2 presents the multilevel model where language scores are controlled for IQ. The posterior means $\hat{\beta}_{0j}^{\text{EB}}$, which can be interpreted as the estimated value added, are presented graphically in Figure 4.4. The figure also presents the confidence intervals for testing the equality of any pair of residuals at a significance level of 5%. The end points of these intervals are given by the posterior mean plus or minus 1.39 times the comparative standard error. For convenience the schools are ordered by the posterior mean.

There are 18 schools having a confidence interval overlapping with the confidence interval of the best school in this sample, implying that their value-added scores do not differ significantly. At the lower extreme, 34 schools do not differ significantly from the worst school in this sample. We can also deduce from the graph that about half of this sample of schools have approximately average scores that cannot be distinguished statistically from the mean value of 0.

Example 4.7 Posterior confidence intervals for random data.

Let us also look, once again, at the random digits example. In Figure 4.5 we have graphed for each of the 10 macro-units (now in the original order) their OLS means and their posterior means with confidence intervals.

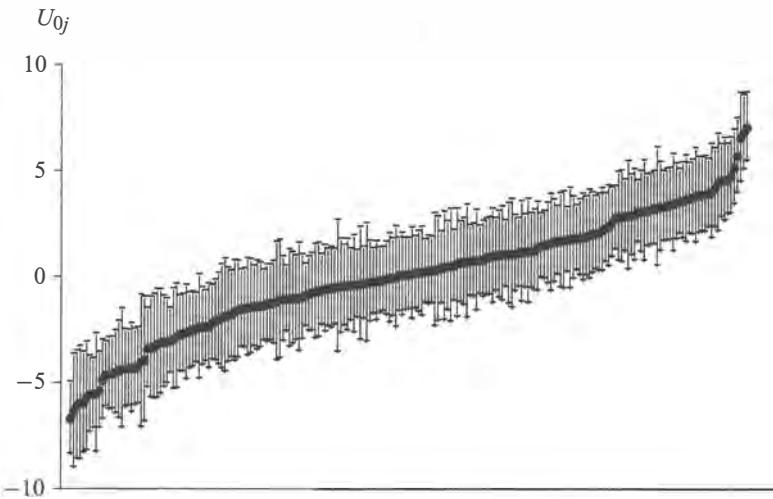


Figure 4.4: The added value scores for 211 schools with comparative posterior confidence intervals.

Once again we clearly observe the shrinkage since the OLS means (\times) are further apart than the posterior means (\bullet). Furthermore, as we would expect from a random digits example, none of the pairwise comparisons results in any significant differences between the macro-units since all 10 confidence intervals overlap.

4.9 Three-level random intercept models

The three-level random intercept model is a straightforward extension of the two-level model. In previous examples, data were used where students were nested within schools. The actual hierarchical structure of educational data is, however, students nested within classes nested within schools. Other examples are: siblings within families within neighborhoods, and people within regions within states. Less obvious examples are: students within cohorts within schools, and longitudinal measurements within persons within groups. These latter cases will be illustrated in Chapter 15 on longitudinal data. For the time being we concentrate on ‘simple’ three-level hierarchical data structures. The dependent variable now is denoted by Y_{ijk} , referring to, for example, pupil i in class j in school k . More generally, one can talk about level-one unit i in level-two unit j in level-three unit k . The three-level model for such data with one explanatory variable may be formulated as a regression model

$$Y_{ijk} = \beta_{0jk} + \beta_1 x_{ijk} + R_{ijk}, \quad (4.20)$$

where β_{0jk} is the intercept in level-two unit j within level-three unit k . For the intercept we have the level-two model,

$$\beta_{0jk} = \delta_{00k} + U_{0jk}, \quad (4.21)$$

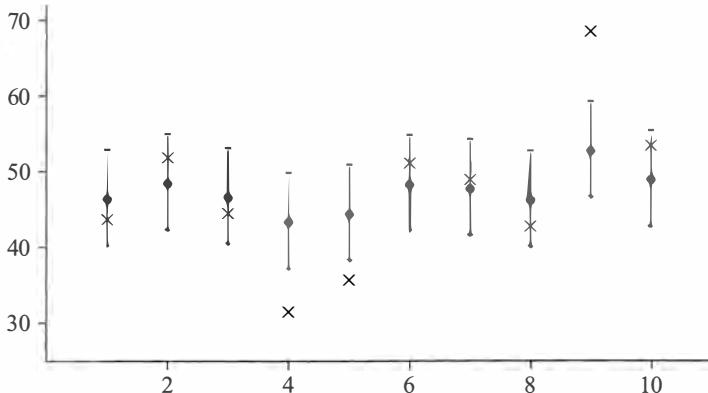


Figure 4.5: OLS means (\times) and posterior means (\bullet) with comparative posterior confidence intervals.

where δ_{00k} is the average intercept in level-three unit k . For this average intercept we have the level-three model,

$$\delta_{00k} = \gamma_{000} + V_{00k}. \quad (4.22)$$

This shows that there are now three residuals, as there is variability on three levels. Their variances are denoted by

$$\text{var}(R_{ijk}) = \sigma^2, \quad \text{var}(U_{0jk}) = \tau^2, \quad \text{var}(V_{00k}) = \varphi^2. \quad (4.23)$$

The total variance between all level-one units now equals $\sigma^2 + \tau^2 + \varphi^2$, and the population variance between the level-two units is $\tau^2 + \varphi^2$. Substituting (4.22) and (4.21) into the level-one model (4.20) and using (in view of the next chapter) the triple indexing notation γ_{100} for the regression coefficient β_1 yields

$$Y_{ijk} = \gamma_{000} + \gamma_{100} x_{ijk} + V_{00k} + U_{0jk} + R_{ijk}. \quad (4.24)$$

Example 4.8 A three-level model: students in classes in schools.

For this example we use a data set on 3,792 students in 280 classes in 57 secondary schools with complete data (see Opdenakker and Van Damme, 1997). On entering the school students were administered tests on IQ, mathematical ability, and achievement motivation, and data were also collected on the educational level of the father and the students' gender.

The response variable is the score on a mathematics test administered at the end of the second grade of secondary school (when the students were approximately 14 years old). Table 4.5 contains the results of the analysis of the empty three-level model (Model 1) and a model with a fixed effect of student intelligence.

The total variance is 11.686, the sum of the three variance components. Since this is a three-level model there are several kinds of intraclass correlation coefficient. Of the total variance, $2.124/11.686 = 18\%$ is situated at the school level while $(2.124 + 1.746)/11.686 = 33\%$ is situated at the class and school level. The level-three intraclass correlation expressing the likeness of

Table 4.5: Estimates for three-level model.

	Model 1		Model 2	
Fixed effects	Coefficient	S.E.	Coefficient	S.E.
γ_{000} = Intercept	7.96	0.23	-4.55	0.50
γ_{100} = Coefficient of IQ			0.121	0.005
Random part	Var. comp.	S.E.	Var. comp.	S.E.
<i>Level-three variance:</i>				
$\varphi_0^2 = \text{var}(V_{00k})$	2.124	0.546	1.109	0.287
<i>Level-two variance:</i>				
$\tau_0^2 = \text{var}(U_{0jk})$	1.746	0.226	0.701	0.116
<i>Level-one variance:</i>				
$\sigma^2 = \text{var}(R_{ijk})$	7.816	0.186	6.910	0.165
Deviance	19,009.7		18,402.7	

students in the same schools thus is estimated to be 0.18, while the intraclass correlation expressing the likeness of students in the same classes and the same schools thus is estimated to be 0.33. In addition, one can estimate the intraclass correlation that expresses the likeness of classes in the same schools. This level-two intraclass correlation is estimated to be $2.124/(2.124 + 1.746) = 0.55$. This is more than 0.5: the school level contributes slightly more to variability than the class level. The interpretation is that if one randomly takes two classes within one school and calculates the average mathematics achievement level in one of the two, one can predict reasonably accurately the average achievement level in the other class. Of course we could have estimated a two-level model as well, ignoring the class level, but that would have led to a redistribution of the class-level variance to the two other levels, and it would affect the validity of hypothesis tests for added fixed effects.

Model 2 shows that the fixed effect of IQ is very strong, with a *t*-ratio (see Section 6.1) of $0.121/0.005 = 24.2$. (The intercept changes drastically because the IQ score does not have a zero mean; the conventional IQ scale, with a population mean of 100, was used.) Adding the effect of IQ leads to a stronger decrease in the class- and school-level variances than in the student-level variance. This suggests that schools and classes are rather internally homogeneous with respect to IQ and/or that intelligence may play its role partly at the school and class levels.

Another example in which the various different types of intraclass correlation coefficient in a three-level model are discussed is Siddiqui et al. (1996).

As in the two-level model, predictor variables at any of the three levels can be added. All features of the two-level model can be generalized quite straightforwardly to the three-level model – significance testing, model building, testing the model fit, centering of variables, etc. – although the researcher should now be more careful because of the more complicated formulation.

For example, for a level-one explanatory variable there may be three kinds of regressions. In the school example, these are the within-class regression, the within-school/between-class regression, and the between-school regression. Coefficients for these

distinct regressions can be obtained by using the class means as well as the school means as explanatory variables with fixed effects.

Example 4.9 Within-class, between-class, and between-school regressions.

Continuing the previous example, we now investigate whether indeed the effect of IQ is in part a class-level or school-level effect: in other words, whether the within-class, between-class/within-school, and between-school regressions are different. Table 4.6 presents the results.

In Model 3, the effects of the class mean, \bar{IQ}_{jk} , as well as the school mean $\bar{\bar{IQ}}_{..k}$, have been added. The class mean has a clearly significant effect ($t = 0.106/0.013 = 8.15$), which indicates that between-class regressions are different from within-class regressions. The school mean does not have a significant effect ($t = 0.039/0.028 = 1.39$), so there is no evidence that the between-school regressions are different from the between-class regressions. It can be concluded that for the explanation of mathematics achievement, the composition with respect to intelligence plays a role at the class level, but not at the school level.

Table 4.6: Estimates for three-level model with distinct within-class, within-school, and between-school regressions.

	Model 3		Model 4	
Fixed effects	Coefficient	S.E.	Coefficient	S.E.
Intercept	-18.16	2.66	-18.14	2.66
Coefficient of IQ_{ijk}	0.107	0.005		
Coefficient of $IQ_{ijk} - \bar{IQ}_{jk}$			0.107	0.005
Coefficient of \bar{IQ}_{jk}	0.106	0.013		
Coefficient of $\bar{IQ}_{jk} - \bar{\bar{IQ}}_{..k}$			0.212	0.012
Coefficient of $\bar{\bar{IQ}}_{..k}$	0.039	0.028	0.252	0.025
Random part	Var. comp.	S.E.	Var. comp.	S.E.
<i>Level-three variance:</i>				
$\text{var}(V_{00k})$	0.798	0.211	0.798	0.211
<i>Level-two variance:</i>				
$\text{var}(U_{0jk})$	0.433	0.089	0.433	0.089
<i>Level-one variance:</i>				
$\text{var}(R_{ijk})$	6.893	0.164	6.893	0.164
Deviance	18,324.3		18,324.3	

As in Section 4.6, replacing the variables by the deviation scores leads to an equivalent model formulation in which, however, the within-class, between-class, and between-school regression coefficients are given directly by the fixed parameters. In the three-level case, this means that we must use the following three variables:

$IQ_{ijk} - \bar{IQ}_{jk}$, the within-class deviation score of the student
from the class mean;

$\bar{IQ}_{jk} - \bar{\bar{IQ}}_{..k}$, the within-school deviation score of the class mean

from the school mean; and
 $\overline{\text{IQ}}_{..k}$, the school mean itself.

The results are shown as Model 4.

We see here that the within-class regression coefficient is 0.107, equal to the coefficient of student-level IQ in Model 3; the between-class/within-school regression coefficient is 0.212, equal (up to rounding errors) to the sum of the student-level and the class-level coefficients in Model 3; while the between-school regression coefficient is 0.252, equal to the sum of all three coefficients in Model 3. From Model 3 we know that the difference between the last two coefficients is not significant.

4.10 Glommary

Regression model. A statistical model for investigating how a *dependent variable* can be predicted, or explained, from one or more *explanatory variables*, also called *independent variables* or *predictor variables*. The usual form for this dependence is a *linear model*.

Intercept. The constant term in a linear function. For the linear function $Y = a + bX$, the intercept is a ; this is the value of Y corresponding to $X = 0$.

Fixed effects. Coefficients in a regression model considered as fixed (i.e., nonrandom) parameters.

Random effects. Coefficients in a regression model considered as random parameters, that is, parameters distributed according to some probability distribution. Normal distributions are often used here, but other distributional shapes can also be used.

Residuals. Another term used for random variables in a regression model, sometimes called ‘errors’, and representing unexplained variability. Random coefficients are a special case. The term R_{ij} in (4.8) is a residual which is not usually called a random coefficient (because coefficients are numbers multiplying something else).

Homoscedasticity. The assumption that residuals for different cases have the same variance. The converse is called heteroscedasticity.

Fixed effects model. A statistical model with fixed effects only, where the only random term is the residual term at the lowest level.

Analysis of covariance model. A linear model with fixed effects, having continuous (interval level) as well as categorical predictor variables.

Ordinary least squares (OLS). The usual way to estimate parameters (fixed effects) in a linear fixed effects model.

Mixed model. A statistical model with fixed as well as random effects.

Intercept. In a linear function of x , the function value at $x = 0$.

Random intercept model. The model to which this chapter is devoted, which is a mixed model with (for two-level structures) two kinds of random residuals: the residual at level one and the random intercept at level two. A general formula is given by (4.8). The three-level random intercept model has three kinds of random residuals, etc.

Hierarchical linear model. A more general model than the random intercept model, and treated in Chapter 5.

Random intercept. The intercept in a random intercept model (also used more generally in hierarchical linear models), symbolized in formula (4.8) by $\gamma_{00} + \gamma_{01}z_{1j} + \dots + \gamma_{0q}z_{qj} + U_{0j}$. Sometimes also the values z are assumed equal to zero for this definition, leading to $\gamma_{00} + U_{0j}$. The random intercept is associated with the *intercept variance*, the variance of U_{0j} .

Empty model. The random intercept model without any explanatory variables, defined by (4.6).

Intraclass correlation coefficient. The correlation between two randomly drawn individuals in the same randomly drawn group; an equivalent definition is the fraction of total variability that is due to the group level. This is defined in the empty model by (4.7). See also Section 3.3.

Within- and between-group regression coefficients. To avoid ecological fallacies, it is important to distinguish within-group from between-group regressions. The within-group regression coefficient of Y on X is the expected difference in Y between two cases in the same group, for a one-unit difference in X . The between-group regression coefficient of Y on X is the expected difference in the group means on Y between two groups differing by one unit in their mean values on X . In Section 4.6 it was explained how these two coefficients can be obtained from the random intercept model in which the individual variable X as well as the group means of X are included as two separate predictor variables.

Cross-level interaction effect. An interaction between two variables defined at different levels.

Parameter estimation. A brief intuitive overview of parameter estimation in the random intercept model was given in Section 4.7.

Maximum likelihood estimation (ML). The best-known general statistical method for parameter estimation, due to R.A. Fisher (1890–1962).

Restricted or residual maximum likelihood estimation (REML). A method of parameter estimation in linear and generalized linear models that often works better than ML, especially for smaller samples, but gives roughly the same results for large sample sizes.

Empirical Bayes estimation. Estimation of the higher-level residuals, such as random intercepts for the groups, by combining group-level and population-level information. This is done using Bayesian techniques, by so-called *posterior means*. The use of population-level information implies a ‘shrinkage’ of the group-level coefficients

toward the population mean, the shrinkage being stronger accordingly as the groups are smaller or, more generally, convey less information.

Posterior confidence intervals. Confidence intervals for higher-level residuals, likewise combining group-level information with population-level information.

Comparative standard error, or posterior standard deviation. The standard error of the empirical Bayes estimate for a group-level residual minus the true (unobserved) residual. This is used for comparing groups.

Diagnostic standard error. The standard error of the empirical Bayes estimate for a group-level residual. This is used for standardizing empirical Bayes estimates for diagnostic model checking.

Higher-level random intercept models. The random intercept model can also be defined for three- and higher-level data structures, and for the three-level case this was discussed in Section 4.9.

5

The Hierarchical Linear Model

In the previous chapter the simpler case of the hierarchical linear model was treated, in which only intercepts are assumed to be random. In the more general case, slopes may also be random. In a study of students within schools, for example, the effect of the pupil's intelligence or socio-economic status on scholastic performance could differ between schools. This chapter presents the general hierarchical linear model, which allows intercepts as well as slopes to vary randomly. The chapter follows the approach of the previous one: most attention is paid to the case of a two-level nesting structure, and the level-one units are called – for convenience only – ‘individuals’, while the level-two units are called ‘groups’. The notation is also the same.

OVERVIEW OF THE CHAPTER

First the concept of random slopes is defined, and a discussion is given of their interpretation and the interpretation of the associated statistical parameters. This involves the notion that random slopes, as well as random intercepts, are group-level latent variables that can be explained by observed group-level variables: ‘intercepts and slopes as outcomes’. Next we discuss how to specify random slope models, in particular, the issue of centering variables by subtracting the overall mean or the group mean. The chapter concludes with brief treatments of parameter estimation and random slopes in three-level models.

5.1 Random slopes

In the random intercept model of Chapter 4, the groups differ with respect to the average value of the dependent variable: the only random group effect is the random intercept. But the relation between explanatory and dependent variables can differ between groups in more ways. For example, in the field of education (nesting structure: students within classrooms), it is possible that the effect of socio-economic status of students on their scholastic achievement is stronger in some classrooms than in others. As an example in developmental psychology (repeated measurements within individual subjects), it is possible that some subjects progress faster than others. In the analysis of covariance, this phenomenon is

known as heterogeneity of regressions across groups, or as group-by-covariate interaction. In the hierarchical linear model, it is modeled by *random slopes*.

Let us go back to a model with group-specific regressions of Y on one level-one variable X only, like model (4.3) but without the effect of Z ,

$$Y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + R_{ij}. \quad (5.1)$$

The intercepts β_{0j} as well as the regression coefficients, or slopes, β_{1j} are group-dependent. These group-dependent coefficients can be split into an average coefficient and the group-dependent deviation:

$$\beta_{0j} = \gamma_{00} + U_{0j}, \quad \beta_{1j} = \gamma_{10} + U_{1j}. \quad (5.2)$$

Substitution leads to the model

$$Y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + U_{0j} + U_{1j}x_{ij} + R_{ij}. \quad (5.3)$$

It is assumed here that the level-two residuals U_{0j} and U_{1j} as well as the level-one residuals R_{ij} have mean 0, given the values of the explanatory variable X . Thus, γ_{10} is the average regression coefficient just as γ_{00} is the average intercept. The first part of (5.3), $\gamma_{00} + \gamma_{10}x_{ij}$, is called the *fixed part* of the model. The second part, $U_{0j} + U_{1j}x_{ij} + R_{ij}$, is called the *random part*.

The term $U_{1j}x_{ij}$ can be regarded as a *random interaction between group and X*. This model implies that the groups are characterized by two random effects: their intercept and their slope. These are called *latent variables*, meaning that they are not directly observed but play a role ‘behind the scenes’ in producing the observed variables. We say that X has a random slope, or a random effect, or a random coefficient. These two group effects will usually not be independent, but correlated. It is assumed that, for different groups, the pairs of random effects (U_{0j}, U_{1j}) are independent and identically distributed, that they are independent of the level-one residuals R_{ij} , and that all R_{ij} are independent and identically distributed. The variance of the level-one residuals R_{ij} is again denoted σ^2 ; the variances and covariance of the level-two residuals (U_{0j}, U_{1j}) are denoted as follows:

$$\begin{aligned} \text{var}(U_{0j}) &= \tau_{00} = \tau_0^2, \\ \text{var}(U_{1j}) &= \tau_{11} = \tau_1^2, \\ \text{cov}(U_{0j}, U_{1j}) &= \tau_{01}. \end{aligned} \quad (5.4)$$

Just as in the preceding chapter, one can say that the unexplained group effects are assumed to be exchangeable.

5.1.1 Heteroscedasticity

Model (5.3) implies not only that individuals within the same group have correlated Y -values (recall the residual intraclass correlation coefficient of Chapter 4), but also that this correlation as well as the variance of Y are dependent on the value of X . For example, suppose that, in a study of the effect of socio-economic status (SES) on scholastic performance (Y), we have schools which do not differ in their effect on high-SES children, but do differ in the effect of SES on Y (e.g., because of teacher expectancy effects). Then for

children from a high-SES background it does not matter which school they go to, but for children from a low-SES background it does. The school then adds a component of variance for the low-SES children, but not for the high-SES children: as a consequence, the variance of Y (for a random child at a random school) will be larger for the former than for the latter children. Further, the intraclass correlation will be nil for high-SES children, whereas for low-SES children it will be positive.

This example shows that model (5.3) implies that the variance of Y , given the value x on X , depends on x . This is called *heteroscedasticity* in the statistical literature. An expression for the variance of (5.3) is obtained as the sum of the variances of the random variables involved plus a term depending on the covariance between U_{0j} and U_{1j} (the other random variables are uncorrelated). Here we also use the independence between the level-one residual R_{ij} and the level-two residuals (U_{0j}, U_{1j}). From (5.3) and (5.4), we obtain the result

$$\text{var}(Y_{ij} | x_{ij}) = \tau_0^2 + 2\tau_{01}x_{ij} + \tau_1^2x_{ij}^2 + \sigma^2. \quad (5.5)$$

Similarly, for two different individuals (i and i' , with $i \neq i'$) in the same group,

$$\text{cov}(Y_{ij}, Y_{i'j} | x_{ij}, x_{i'j}) = \tau_0^2 + \tau_{01}(x_{ij} + x_{i'j}) + \tau_1^2 x_{ij} x_{i'j}, \quad (5.6)$$

Formula (5.5) implies that the residual variance of Y is minimal for $x_{ij} = -\tau_{01}/\tau_{11}$. (This is deduced by differentiation with respect to x_{ij} .) When this value is within the range of possible X -values, the residual variance first decreases and then increases again; if this value is smaller than all X -values, then the residual variance is an increasing function of x ; if it is larger than all X -values, then the residual variance is decreasing.

5.1.2 Do not force τ_{01} to be 0!

All of the preceding discussion implies that the group effects depend on x : according to (5.3), this effect is given by $U_{0j} + U_{1j}x$. This is illustrated by Figure 5.1, a hypothetical graph of the regression of school achievement (Y) on intelligence (X).

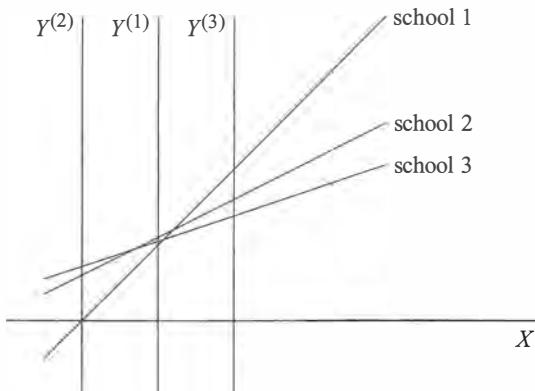


Figure 5.1: Different vertical axes.

It is clear that there are slope differences between the three schools. Looking at the $Y^{(1)}$ -axis, there are almost no intercept differences between the schools. But if we add a value 10 to each intelligence score x , then the Y -axis is shifted to the left by 10 units: the $Y^{(2)}$ -axis. Now school 3 is the best, school 1 the worst: there are strong intercept differences. If we had subtracted 10 from the x -scores, we would have obtained the $Y^{(3)}$ -axis, again with intercept differences but now in reverse order. This implies that the intercept variance τ_{00} , as well as the intercept-by-slope covariance τ_{01} , depend on the origin (0-value) for the X -variable. From this we can learn two things:

1. Since the origin of most variables in the social sciences is arbitrary, in random slope models the intercept-by-slope covariance should be a free parameter estimated from the data, and not *a priori* constrained to the value 0 (i.e., left out of the model).
2. In random slope models we should be careful with the interpretation of the intercept variance and the intercept-by-slope covariance, since the intercept refers to an individual with $x = 0$. For the interpretation of these parameters it is helpful to define the scale for X so that $x = 0$ has an interpretable meaning, preferably as a reference situation. For example, in repeated measurements when X refers to time, or measurement number, it can be convenient to let $x = 0$ correspond to the start, or the end, of the measurements. In nesting structures of individuals within groups, it is often convenient to let $x = 0$ correspond to the overall mean of the population or the sample – for example, if X is IQ at the conventional scale with mean 100, it is advisable to subtract 100 to obtain a population mean of 0.

5.1.3 Interpretation of random slope variances

For the interpretation of the variance of the random slopes, τ_1^2 , it is also illuminating to take the average slope, γ_{10} , into consideration. Model (5.3) implies that the regression coefficient, or slope, for group j is $\gamma_{10} + U_{1j}$. This is a normally distributed random variable with mean γ_{10} and standard deviation $\tau_1 = \sqrt{\tau_1^2}$. Since about 95% of the probability of a normal distribution is within two standard deviations from the mean, it follows that approximately 95% of the groups have slopes between $\gamma_{10} - 2\tau_1$ and $\gamma_{10} + 2\tau_1$. Conversely, about 2.5% of the groups have a slope less than $\gamma_{10} - 2\tau_1$ and 2.5% have a slope steeper than $\gamma_{10} + 2\tau_1$.

Example 5.1 A random slope for IQ.

We continue our study of the effect of IQ on a language test score, begun in Chapter 4. Recall that IQ is here on a scale with mean 0 and its standard deviation in this data set is 2.04. A random slope of IQ is added to the model, that is, the effect of IQ is allowed to differ between classes. The model is an extension of model (5.3): a fixed effect for the class average on IQ is added. The model reads

$$Y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + \gamma_{01}\bar{x}_j + U_{0j} + U_{1j}x_{ij} + R_{ij}.$$

The results can be read from Table 5.1. Note that the ‘Level-two random part’ heading refers to the random intercept and random slope which are random effects associated with the level-two units (the class), but that the variable that has the random slope, IQ, is itself a level-one variable.

Table 5.1: Estimates for random slope model.

Fixed effect	Coefficient	S.E.
γ_{00} = Intercept	41.127	0.234
γ_{10} = Coefficient of IQ	2.480	0.064
γ_{01} = Coefficient of \bar{IQ} (group mean)	1.029	0.262
Random part	Parameters	S.E.
<i>Level-two random part:</i>		
$\tau_0^2 = \text{var}(U_{0j})$	8.877	1.117
$\tau_1^2 = \text{var}(U_{1j})$	0.195	0.076
$\tau_{01} = \text{cov}(U_{0j}, U_{1j})$	-0.835	0.217
<i>Level-one variance:</i>		
$\sigma^2 = \text{var}(R_{ij})$	39.685	0.964
Deviance	24,864.9	

Figure 5.2 presents a sample of 15 regression lines, randomly chosen according to the model of Table 5.1. (The values of the group mean \bar{IQ} were chosen randomly from a normal distribution with mean -0.072 and standard deviation 0.955 , which are the mean and standard deviation of the group mean of IQ in this data set.) This figure thus demonstrates the population of regression lines that characterizes, according to this model, the population of schools.

Should the value of 0.195 for the random slope variance be considered to be high? The slope standard deviation is $\sqrt{0.195} = 0.44$, and the average slope is $\gamma_{10} = 2.48$. The values of average slope

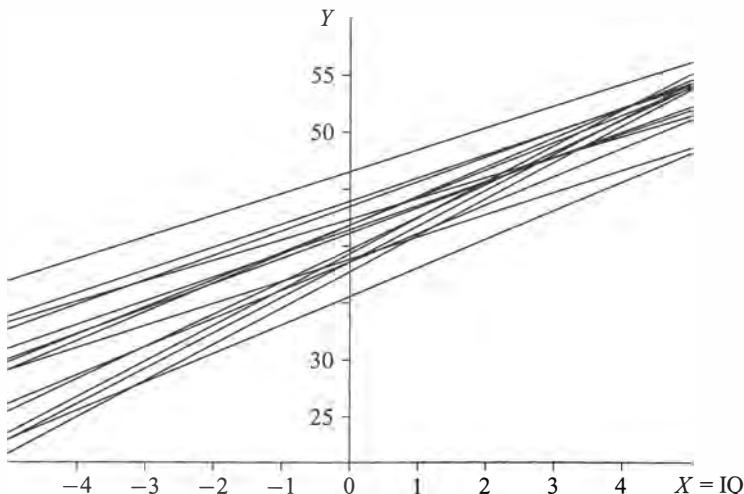


Figure 5.2: Fifteen random regression lines according to the model of Table 5.1 (with randomly chosen intercepts and slopes).

\pm two standard deviations range from 1.60 to 3.36. This implies that the effect of IQ is clearly positive in all classes, but high effects of IQ are more than twice as large as low effects. This may indeed be considered an important difference. (As indicated above, ‘high’ and ‘low’ are respectively understood here as those values occurring in classes with the top 2.5% and the bottom 2.5% of the class-dependent effects.)

Interpretation of intercept-slope covariance

The correlation between random slope and random intercept is $-0.83/\sqrt{8.88 \times 0.195} = -0.63$. Recall that all variables are centered (have zero mean), so that the intercept corresponds to the language test score for a pupil with average intelligence in a class with average mean intelligence. The negative correlation between slope and intercept means that classes with a higher performance for a pupil of average intelligence have a lower within-class effect of intelligence. Thus, the higher average performance tends to be achieved more by higher language scores of the less intelligent, than by higher scores of the more intelligent students.

In a random slope model, the within-group coherence cannot be simply expressed by the intra-class correlation coefficient or its residual version. The reason is that, in terms of the present example, the correlation between students in the same class depends on their intelligence. Thus, the extent to which a given classroom deserves to be called ‘good’ varies across students.

To investigate how the contribution of classrooms to students’ performance depends on IQ, consider the equation implied by the parameter estimates:

$$Y_{ij} = 41.13 + 2.480 \text{IQ}_{ij} + 1.029 \bar{\text{IQ}}_j + U_{0j} + U_{1j} \text{IQ}_{ij} + R_{ij}.$$

Recall from Example 4.2 that the standard deviation of the IQ score is about 2, and the mean is 0. Hence students with an intelligence among the bottom few percent or the top few percent have IQ scores of about ± 4 . Substituting these values in the contribution of the random effects gives $U_{0j} \pm 4U_{1j}$. It follows from equations (5.5) and (5.6) that for students with $\text{IQ} = \pm 4$, we have

$$\begin{aligned}\text{var}(Y_{ij} | \text{IQ}_{ij} = -4) &= 8.88 + 2 \times (-0.835) \times (-4) + (-4)^2 \times 0.195 + 39.69 = 58.37, \\ \text{cov}(Y_{ij}, Y_{i'j} | \text{IQ}_{ij} = -4, \text{IQ}_{i'j} = 4) &= 8.88 - 16 \times 0.195 = 5.76, \\ \text{var}(Y_{ij} | \text{IQ}_{ij} = 4) &= 8.88 - 8 \times 0.835 + 16 \times 0.195 + 39.69 = 45.01,\end{aligned}$$

and therefore

$$\rho(Y_{ij}, Y_{i'j} | \text{IQ}_{ij} = -4, \text{IQ}_{i'j} = 4) = \frac{5.76}{\sqrt{58.37 \times 45.01}} = 0.11.$$

Hence, the language test scores of the most intelligent and the least intelligent students in the same class are positively correlated over the population of classes: classes that have relatively good results for the less able tend also to have relatively good results for the more able students.

This positive correlation corresponds to the result that the value of IQ for which the variance given by (5.5) is minimal, is outside the range from -4 to $+4$. For the estimates in Table 5.1, this variance is (with some rounding)

$$\text{var}(Y_{ij} | \text{IQ}_{ij} = x) = 8.88 - 1.67x + 0.195x^2 + \sigma^2.$$

Taking the derivative of this function of x equating it to 0 yields that the variance is minimal for $x = 1.67/0.39 = 4.3$, just outside the IQ range from -4 to $+4$. This again implies that classes tend mostly to perform either higher, or lower, over the entire range of IQ. This is illustrated also by Figure 5.2 (which, however, also contains some regression lines that cross each other within the range of IQ, illustrating that the random nature of these regression lines will lead to exceptions to this pattern).

5.2 Explanation of random intercepts and slopes

The aim of regression analysis is to explain variability in the outcome (i.e., dependent) variable. Explanation is understood here in a quite limited way – as being able to predict the value of the dependent variable from knowledge of the values of the explanatory variables. The unexplained variability in single-level multiple regression analysis is just the variance of the residual term. Variability in multilevel data, however, has a more complicated structure. This is related to the fact, mentioned in the preceding chapter, that several populations are involved in multilevel modeling: one population for each level. Explaining variability in a multilevel structure can be achieved by explaining variability between individuals but also by explaining variability between groups; if there are random slopes as well as random intercepts, at the group level one could try to explain the variability of slopes as well as intercepts.

In the model defined by (5.1)–(5.3), some variability in Y is explained by the regression on X , that is, by the term $\gamma_{10} x_{ij}$; the random coefficients U_{0j} , U_{1j} , and R_{ij} each express different parts of the unexplained variability. In order to try to explain more of the unexplained variability, all three of these can be the point of attack. In the first place, one can try to find explanations in the population of individuals (at level one). The part of residual variance that is expressed by $\sigma^2 = \text{var}(R_{ij})$ can be diminished by including other level-one variables. Since group compositions with respect to level-one variables can differ from group to group, inclusion of such variables may also diminish residual variance at the group level. A second possibility is to try to find explanations in the population of groups (at level two). Note that here the variables that we try to explain, U_{0j} and U_{1j} , are not directly observed, but latent variables. If we wish to reduce the unexplained variability associated with U_{0j} and U_{1j} , we can also say that we wish to expand equations (5.2) by predicting the group-dependent regression coefficients β_{0j} and β_{1j} from level-two variables Z . Supposing for the moment that we have one such variable, this leads to regression formulas for β_{0j} and β_{1j} on the variable Z given by

$$\beta_{0j} = \gamma_{00} + \gamma_{01} z_j + U_{0j} \quad (5.7)$$

and

$$\beta_{1j} = \gamma_{10} + \gamma_{11} z_j + U_{1j}. \quad (5.8)$$

In words, the β s are treated as dependent variables in regression models for the population of groups; however, these are ‘latent regressions’, because the β s cannot be observed without error. Equation (5.7) is called an *intercepts as outcomes* model, and (5.8) a *slopes as outcomes* model.¹

¹In the older literature, these equations were applied to the estimated groupwise regression coefficients rather than the latent coefficients. The statistical estimation then was carried out in two stages: first ordinary least squares (OLS) estimation within each group, then OLS estimation with the estimated coefficients as outcomes. This is statistically inefficient unless the group sizes are very large, and does not distinguish the ‘true score’ variability of the latent coefficients from the sampling variability of the estimated groupwise regression coefficients. A two-step approach which does make this distinction is briefly treated in Section 3.7, following equation (3.40).

5.2.1 Cross-level interaction effects

This chapter began with the basic model (5.1), reading

$$Y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + R_{ij}.$$

Substituting (5.7) and (5.8) into this equation leads to the model

$$\begin{aligned} Y_{ij} &= (\gamma_{00} + \gamma_{01}z_j + U_{0j}) + (\gamma_{10} + \gamma_{11}z_j + U_{1j})x_{ij} + R_{ij} \\ &= \gamma_{00} + \gamma_{01}z_j + \gamma_{10}x_{ij} + \gamma_{11}z_jx_{ij} \\ &\quad + U_{0j} + U_{1j}x_{ij} + R_{ij}, \end{aligned} \tag{5.9}$$

which we have rearranged so that the fixed part comes first and the random part next. Comparing this with model (5.3) shows that this explanation of the random intercept and slope leads to a different fixed part of the model, but does not change the formula for the random part, which remains $U_{0j} + U_{1j}x_{ij} + R_{ij}$. However, it is to be expected that the residual random intercept and slope variances, τ_0^2 and τ_1^2 , will be less than their counterparts in model (5.3) because part of the variability of intercept and slopes now is explained by Z . In Chapter 7 we will see, however, that this is not necessarily so for the estimated values of these parameters.

Equation (5.9) shows that explaining the intercept β_{0j} by a level-two variable Z leads to a main effect of Z , while explaining the coefficient β_{1j} of X by the level-two variable Z leads to a product interaction effect of X and Z . Such an interaction between a level-one and a level-two variable is called a *cross-level interaction*.

For the definition of interaction variables such as the product z_jx_{ij} in (5.9), it is advisable to use component variables Z and X for which the values $Z = 0$ and $X = 0$, respectively, have some interpretable meaning. For example, the variables Z and X could be centered around their means, so that $Z = 0$ means that Z has its average value, and analogously for X . Another possibility is that the zero values correspond to some kind of reference value. The reason is that, in the presence of the interaction term $\gamma_{11}z_jx_{ij}$, the main effect coefficient γ_{10} of X is to be interpreted as the effect of X for cases with $Z = 0$, while the main effect coefficient γ_{01} of Z is to be interpreted as the effect of Z for cases with $X = 0$.

Example 5.2 Cross-level interaction between IQ and group-mean IQ.

A comparison of Examples 4.2 and 4.3 shows that part of the intercept variability in the language scores is explained by the class mean of IQ. This confirms that this aggregated variable is a meaningful classroom characteristic. We attempt to use it also to explain part of the random slope variability, which means that we introduce the interaction between individual IQ and the classroom mean of IQ. When this variable is added to the model of Example 5.1, the parameter estimates presented in Table 5.2 are obtained.

The cross-level interaction is significant (tested by a t -ratio as explained in Section 6.1), with $t = -0.187/0.064 = -2.92$, $p < 0.01$, and the random slope variance has decreased from 0.195 to 0.163. (It follows from Chapter 7, however, that the expectation that a random slope variance should decrease when a cross-level interaction effect is included, is not always confirmed.) Thus, the classroom average of IQ is indeed successful in explaining part of the between-group variability in the effect of IQ.

How can we assess the magnitude of this cross-level effect? The class-dependent regression coefficient of IQ (cf. (5.8)) is

$$\gamma_{10} + \gamma_{12}z_j + U_{1j},$$

Table 5.2: Estimates for model with random slope and cross-level interaction.

Fixed effect	Coefficient	S.E.
γ_{00} = Intercept	41.254	0.235
γ_{10} = Coefficient of IQ	2.463	0.063
γ_{01} = Coefficient of \bar{IQ}	1.131	0.262
γ_{11} = Coefficient of $\bar{IQ} \times IQ$	-0.187	0.064
Random part	Parameters	S.E.
<i>Level-two random part:</i>		
$\tau_0^2 = \text{var}(U_{0j})$	8.601	1.088
$\tau_1^2 = \text{var}(U_{1j})$	0.163	0.072
$\tau_{01} = \text{cov}(U_{0j}, U_{1j})$	-0.833	0.210
<i>Level-one variance:</i>		
$\sigma^2 = \text{var}(R_{ij})$	39.758	0.965
Deviance	24,856.8	

estimated as

$$2.463 - 0.187 \bar{IQ} + U_{1j} .$$

The classroom mean of the centered IQ variable has a rather skewed distribution, ranging from -4.8 to +2.5. For \bar{IQ} ranging between -4.8 and +2.5, the fixed part of this expression ranges (in reverse order) between about 2.00 and 3.36. This implies sizeable differences in the effect of intelligence on language score between classes with students having low average IQ and those with students of high average IQ, classrooms with lower average IQ being associated with higher effects of intelligence on language scores. (We must note, however, that further analysis of these data will show that the final conclusions will be different.)

Cross-level interactions can be considered on the basis of two different kinds of argument. The above presentation is in line with an inductive argument: a researcher who finds a significant random slope variance may be led to think of level-two variables that could explain the random slope. An alternative approach is to base the cross-level interaction on substantive (theoretical) arguments formulated before looking at the data. The researcher is then led to estimate and test the cross-level interaction effect irrespective of whether a random slope variance was found. If a cross-level interaction effect exists, the power of the statistical test of this fixed effect is considerably higher than the power of the test for the corresponding random slope (assuming that the same model serves as the null hypothesis). Therefore there is nothing wrong with looking for a specific cross-level interaction even if no significant random slope was found. This is further discussed in the last part of Section 6.4.1.

More variables

The preceding models can be extended by including more variables that have random effects, and more variables explaining these random effects. Suppose that there are p level-one explanatory variables, X_1, \dots, X_p , and q level-two explanatory variables, Z_1, \dots, Z_q .

Then, if the researcher is not afraid of a model with too many parameters, he can consider the model where all X -variables have varying slopes, and where the random intercept as well as all these slopes are explained by all Z -variables. At the within-group level (i.e., for the individuals) the model is then a regression model with p variables,

$$Y_{ij} = \beta_{0j} + \beta_{1j} x_{1ij} + \cdots + \beta_{pj} x_{pij} + R_{ij}. \quad (5.10)$$

The explanation of the regression coefficients $\beta_{0j}, \beta_{1j}, \dots, \beta_{pj}$ is based on the between-group model, which is a q -variable regression model for the group-dependent coefficient β_{hj} ,

$$\beta_{hj} = \gamma_{h0} + \gamma_{h1} z_{1j} + \cdots + \gamma_{hq} z_{qj} + U_{hj}. \quad (5.11)$$

Substituting (5.11) into (5.10) and rearranging terms then yields the model

$$\begin{aligned} Y_{ij} = & \gamma_{00} + \sum_{h=1}^p \gamma_{h0} x_{hij} + \sum_{k=1}^q \gamma_{0k} z_{kj} + \sum_{k=1}^q \sum_{h=1}^p \gamma_{hk} z_{kj} x_{hij} \\ & + U_{0j} + \sum_{h=1}^p U_{hj} x_{hij} + R_{ij}. \end{aligned} \quad (5.12)$$

This shows that we obtain main effects of each X and Z variable as well as all cross-level product interactions. Further, we see the reason why, in formula (4.8), the fixed coefficients were called γ_{h0} for the level-one variable X_h and γ_{0k} for the level-two variable Z_k .

The groups are now characterized by $p+1$ random coefficients $U_{0j}, U_{1j}, \dots, U_{pj}$. These random coefficients are independent between groups, but may be correlated within groups. It is assumed that the vector (U_{0j}, \dots, U_{pj}) is independent of the level-one residuals R_{ij} and that all residuals have population mean 0, given the values of all explanatory variables. It is also assumed that the level-one residual R_{ij} has a normal distribution with constant variance σ^2 and that (U_{0j}, \dots, U_{pj}) has a multivariate normal distribution with a constant covariance matrix. Analogously to (5.4), the variances and covariances of the level-two random effects are denoted by

$$\begin{aligned} \text{var}(U_{hj}) &= \tau_{hh} = \tau_h^2 \quad (h = 1, \dots, p), \\ \text{cov}(U_{hj}, U_{kj}) &= \tau_{hk} \quad (h, k = 1, \dots, p). \end{aligned} \quad (5.13)$$

Interactions and ecological fallacies

In Section 3.1 we saw that aggregations of level-one variables to level two (or higher levels) can be substantively meaningful and can be important to avoid ecological fallacies. This was discussed in Section 4.6, showing how the group-level average of an individual-level variable can be included in the model to represent the difference between within-group and between-group regressions. Similar considerations apply to interaction effects. For example, when the interaction effect of two level-one variables X_1 and X_2 is being investigated, we should be aware of the possibility that this interaction has effects at level two rather than at level one, for one or both of the variables. Thus, the representation of the interaction by a product variable $x_{1ij}x_{2ij}$ can be extended to four possibilities,

$$x_{1ij}x_{2ij}, \bar{x}_{1j}x_{2ij}, x_{1ij}\bar{x}_{2j}, \text{ and } \bar{x}_{1j}\bar{x}_{2j}, \quad (5.14)$$

each with their own theoretical interpretation. An example was given by van Yperen and Snijders (2000) in a study of the interaction between job demands and job control on the psychological health of employees in departments of firms.

Example 5.3 A model with many fixed effects.

Next to intelligence, another important explanation of school performance is the home environment. This is reflected by the socio-economic status of the pupil's family, represented here by a variable called SES (a numerical variable with mean 0.0 and standard deviation 10.9 in this data set). We include SES at the individual level as well as the class mean; we are also interested in the interaction effect between IQ and SES. To try and avoid ecological fallacies, in line with the discussion above, the four combinations of interactions between IQ and SES are included as in (5.14).

The covariance of the random slopes of IQ and SES is excluded, because including this covariance in the model led to failure of the estimating algorithm to converge. We may assume that this covariance is not different from 0 to any important extent.

Estimating model (5.12) (with $p = q = 2$), to which are added the interaction between the two individual variables as well as the interaction between the two classroom means, leads to the results presented in Table 5.3. A discussion of this table is deferred to the next example.

Table 5.3: Estimates for model with random slopes and many effects.

Fixed effect	Coefficient	S.E.
γ_{00} = Intercept	41.632	0.255
γ_{10} = Coefficient of IQ	2.230	0.063
γ_{20} = Coefficient of SES	0.172	0.012
γ_{30} = Interaction of IQ and SES	-0.019	0.006
γ_{01} = Coefficient of \bar{IQ}	0.816	0.308
γ_{02} = Coefficient of \bar{SES}	-0.090	0.044
γ_{03} = Interaction of \bar{IQ} and \bar{SES}	-0.134	0.037
γ_{11} = Interaction of IQ and \bar{IQ}	-0.081	0.081
γ_{12} = Interaction of IQ and \bar{SES}	0.004	0.013
γ_{21} = Interaction of SES and \bar{IQ}	0.023	0.018
γ_{22} = Interaction of SES and \bar{SES}	0.000	0.002
Random part	Parameters	S.E.
<i>Level-two random part:</i>		
$\tau_0^2 = \text{var}(U_{0j})$	8.344	1.407
$\tau_1^2 = \text{var}(U_{1j})$	0.165	0.069
$\tau_{01} = \text{cov}(U_{0j}, U_{1j})$	-0.942	0.204
$\tau_2^2 = \text{var}(U_{2j})$	0.0	0.0
$\tau_{02} = \text{cov}(U_{0j}, U_{2j})$	0.0	0.0
<i>Level-one variance:</i>		
$\sigma^2 = \text{var}(R_{ij})$	37.358	0.907
Deviance	24,624.0	

Unless p and q are quite small, model (5.12) entails a number of statistical parameters that is usually too large for comfort. Therefore, two simplifications are often used.

1. Not all X -variables are considered to have random slopes. Note that the explanation of the variable slopes by the Z -variables may have led to some random residual coefficients that are not significantly different from 0 (testing the τ parameters is discussed in Chapter 6) or even estimated to be 0.
2. Given that the coefficients β_{hj} of a certain variable X_h are variable across groups, it is not necessary to use all variables Z_k in explaining their variability. The number of cross-level interactions can be restricted by explaining each β_{hj} by only a well-chosen subset of the Z_k .

Which variables to give random slopes, and which cross-level and other interaction variables to use, will depend on subject-matter as well as empirical considerations. The statistical aspects of testing and model fitting are treated in later chapters. For further help in the interpretation of interactions, in particular, methods to determine the values of z_j where the regression on x_{ij} is different from zero (or vice versa), see Curran and Bauer (2005) and Preacher et al. (2006).

Example 5.4 A parsimonious model with several variables.

In Table 5.3 the random slope variance of SES is estimated by 0 (this happens occasionally; cf. Section 5.4). Therefore, this random slope is excluded from the model.

We shall see in Chapter 6 that the significance of fixed effects can be tested by applying a t -test to the ratio of estimate to standard error. This consideration led us to exclude all cross-level interactions. The resulting estimates are displayed in Table 5.4. The estimates of the remaining effects do not change much compared to Table 5.3.

All explanatory variables have mean 0. Therefore, the intercept γ_{00} is the mean of students with average IQ and SES in a class with average composition with respect to these variables. The regression coefficient of IQ γ_{10} is the average effect of IQ for students of average SES, and similarly for the regression coefficient γ_{20} for SES.

Compared to Example 5.2, it turns out that there is no significant cross-level interaction effect between IQ and its classroom mean, but there are two interaction effects between IQ and SES: the interaction between the two individual variables and the interaction between the two classroom means.

The model found can be expressed as a model with variable intercepts and slopes by the formula

$$Y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + \beta_{2j}x_{2ij} + \beta_{3j}x_{3ij} + R_{ij},$$

where X_1 is IQ, X_2 is SES, and X_3 is their product interaction. The intercept is

$$\beta_{0j} = \gamma_{00} + \gamma_{01}z_{1j} + \gamma_{02}z_{2j} + \gamma_{03}z_{3j} + U_{0j},$$

where Z_1 is average IQ, Z_2 is average SES, and Z_3 is their product interaction. The coefficient of X_1 is

$$\beta_{1j} = \gamma_{10} + U_{1j},$$

while the coefficient of X_2 is not variable,

$$\beta_{2j} = \gamma_{20}.$$

Table 5.4: Estimates for a more parsimonious model.

Fixed effect	Coefficient	S.E.
γ_{00} = Intercept	41.612	0.247
γ_{10} = Coefficient of IQ	2.231	0.063
γ_{20} = Coefficient of SES	0.174	0.012
γ_{30} = Interaction of IQ and SES	-0.017	0.005
γ_{01} = Coefficient of \overline{IQ}	0.760	0.296
γ_{02} = Coefficient of \overline{SES}	-0.089	0.042
γ_{03} = Interaction of \overline{IQ} and \overline{SES}	-0.120	0.033
Random part	Parameters	S.E.
<i>Level-two random part:</i>		
$\tau_0^2 = \text{var}(U_{0j})$	8.369	1.050
$\tau_1^2 = \text{var}(U_{1j})$	0.164	0.069
$\tau_{01} = \text{cov}(U_{0j}, U_{1j})$	-0.929	0.204
<i>Level-one variance:</i>		
$\sigma^2 = \text{var}(R_{ij})$	37.378	0.907
Deviance	24,626.8	

5.2.2 A general formulation of fixed and random parts

Formally, and in many computer programs, these simplifications lead to a representation of the hierarchical linear model that is slightly different from (5.12). (For example, the HLM program uses the formulations (5.10)–(5.12), whereas R and MLwiN use formulation (5.15).) Whether a level-one variable was obtained as a cross-level interaction or not is immaterial to the computer program. Even the difference between level-one variables and level-two variables, although possibly relevant for the way the data are stored, is of no importance for the parameter estimation. Therefore, all variables – level-one and level-two variables, including product interactions – can be represented mathematically simply as x_{hij} . When there are r explanatory variables, ordered so that the first p have fixed *and* random coefficients, while the last $r - p$ have only fixed coefficients,² the hierarchical linear model can be represented as

$$Y_{ij} = \gamma_0 + \sum_{h=1}^p \gamma_h x_{hij} + U_{0j} + \sum_{h=1}^p U_{hj} x_{hij} + R_{ij}. \quad (5.15)$$

The two terms,

$$\gamma_0 + \sum_{h=1}^r \gamma_h x_{hij} \quad \text{and} \quad U_{0j} + \sum_{h=1}^p U_{hj} x_{hij} + R_{ij},$$

²It is mathematically possible that some variables have a random but not a fixed effect. This makes sense only in special cases.

are the fixed and the random parts of the model, respectively.

In cases where the explanation of the random effects works extremely well, one may end up with models with no random effects at level two. In other words, the random intercept U_{0j} and all random slopes U_{hj} in (5.15) have zero variance, and may just as well be omitted from the formula. In this case, the resulting model may be analyzed just as well by OLS regression analysis, because the residuals are independent and have constant variance. Of course, this is known only after the multilevel analysis has been carried out. In such a case, the within-group dependence between measurements has been fully explained by the available explanatory variables (and their interactions). This underlines the fact that whether the hierarchical linear model is a more adequate model for analysis than OLS regression depends not on the mutual dependence of the measurements, but on the mutual dependence of the residuals.

5.3 Specification of random slope models

Given that random slope models are available, the researcher has many options to model his data. Each predictor may be assigned a random slope, and each random slope may covary with any other random slope. Parsimonious models, however, should be preferred, if only for the simple reason that a strong scientific theory is general rather than specific. A good reason for choosing between a fixed and a random slope for a given predictor variable should preferably be found in the theory that is being investigated. If the theory (whether this is a general scientific theory or a practical policy theory) does not give any clue with respect to a random slope for a certain predictor variable, then one may be tempted to refrain from using random slopes. However, this implies a risk of invalid statistical tests, because if some variable does have a random slope, then omitting this feature from the model could affect the estimated standard errors of the other variables. The specification of the hierarchical linear model, including the random part, is discussed more fully in Section 6.4 and Chapter 10.

In data exploration, one can try various specifications. Often it appears that the chance of detecting slope variation is high for variables with strong fixed effects. This, however, is an empirical rather than a theoretical assertion. Actually, it may well be that when a fixed effect is – almost – zero, there does exist slope variation. Consider, for instance, the case where male teachers treat boys advantageously over girls, while female teachers do the reverse. If half of the sample consists of male and the other half of female teachers, then, all other things being equal, the main gender effect on achievement will be absent, since in half of the classes the gender effect will be positive and in the other half negative. The fixed effect of students' gender is then zero but varies across classes (depending on the teachers' gender). In this example, of course, the random effect would disappear if one specified the cross-level interaction effect of teachers' gender with students' gender.

5.3.1 Centering variables with random slopes?

Recall from Figure 5.1 that the intercept variance and the meaning of the intercept in random slope models depend on the location of the X -variable. Also the covariance between the intercepts and the slopes is dependent on this location. In the examples presented so far we have used an IQ score for which the grand mean was zero (the original score was

transformed by subtracting the grand mean IQ). This facilitated interpretation since the intercept could be interpreted as the expected score for a student with average IQ. Making the IQ slope random did not have consequences for these meanings.

In Section 4.6 a model was introduced by which we could distinguish within- from between-group regression. Two models were discussed:

$$Y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + \gamma_{01}\bar{x}_j + U_{0j} + R_{ij} \quad (4.9)$$

and

$$Y_{ij} = \tilde{\gamma}_{00} + \tilde{\gamma}_{10}(x_{ij} - \bar{x}_j) + \tilde{\gamma}_{01}\bar{x}_j + U_{0j} + R_{ij}. \quad (4.10)$$

It was shown that $\tilde{\gamma}_{01} = \gamma_{10} + \gamma_{01}$, $\tilde{\gamma}_{10} = \gamma_{10}$, so that the two models are equivalent.

Are the models also equivalent when the effect of X_{ij} or $(X_{ij} - \bar{X}_j)$ is random across groups? This was discussed by Kreft et al. (1995). Let us first consider the extension of (4.9). Define the level-one and level-two models

$$Y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \gamma_{01}\bar{x}_j + R_{ij},$$

$$\beta_{0j} = \gamma_{00} + U_{0j},$$

$$\beta_{1j} = \gamma_{10} + U_{1j};$$

substituting the level-two model into the level-one model leads to

$$Y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + \gamma_{01}\bar{x}_j + U_{0j} + U_{1j}x_{ij} + R_{ij}.$$

Next we consider the extension of (4.10):

$$Y_{ij} = \tilde{\beta}_{0j} + \tilde{\beta}_{1j}(x_{ij} - \bar{x}_j) + \tilde{\gamma}_{01}\bar{x}_j + R_{ij},$$

$$\tilde{\beta}_{0j} = \tilde{\gamma}_{00} + U_{0j},$$

$$\tilde{\beta}_{1j} = \tilde{\gamma}_{10} + U_{1j};$$

substitution and rearrangement of terms now yields

$$Y_{ij} = \tilde{\gamma}_{00} + \tilde{\gamma}_{10}x_{ij} + (\tilde{\gamma}_{01} - \tilde{\gamma}_{10})\bar{x}_j + U_{0j} + U_{1j}x_{ij} - U_{1j}\bar{x}_j + R_{ij}.$$

This shows that the two models differ in the term $U_{1j}\bar{x}_j$ which is included in the group-mean-centered random slope model but not in the other model. Therefore in general there is no one-to-one relation between the γ and the $\tilde{\gamma}$ parameters, so the models are not statistically equivalent except for the extraordinary case where variable X has no between-group variability.

This implies that in constant slope models one can either use X_{ij} and \bar{X}_j or $(X_{ij} - \bar{X}_j)$ and \bar{X}_j as predictors, since this results in statistically equivalent models, but in random slope models one should carefully choose one or the other specification depending on substantive considerations and/or model fit.

On which consideration should this choice be based? Generally one should be reluctant to use group-mean-centered random slopes models unless there is a clear theory (or an empirical clue) that not the absolute score X_{ij} but rather the relative score $(X_{ij} - \bar{X}_j)$ is related to Y_{ij} . Now $(X_{ij} - \bar{X}_j)$ indicates the relative position of an individual in his or her group, and examples of instances where one may be particularly interested in this variable are:

- * research on normative or comparative social reference processes (e.g., Guldemond, 1994);
- * research on relative deprivation;
- * research on teachers' rating of student performance.

A more detailed discussion is given in Enders and Tofghi (2007). However, that paper pays no attention to the fact that interactions between level-one variables can also play a role for their group means (see our discussion on p. 83). This implies further decisions about centering, which likewise should be based on a combination of substantive theoretical and empirical considerations.

5.4 Estimation

The discussion in Section 4.7 can also be applied, with the necessary extensions, to the estimation of parameters in the more complicated model (5.15). A number of iterative estimation algorithms have been proposed by, for example, Laird and Ware (1982), Goldstein (1986), and Longford (1987), and are now implemented in multilevel software.

The following may give some intuitive understanding of estimation methods. If the parameters of the random part, that is, the parameters in (5.13) together with σ^2 , were known, then the regression coefficients could be estimated straightforwardly with the so-called generalized least squares (GLS) method. Conversely, if all regression coefficients γ_{hk} were known, the ‘total residuals’ (which seems an apt name for the second line of equation (5.12)) could be computed, and their covariance matrix could be used to estimate the parameters of the random part. These two partial estimation processes can be alternated: use provisional values for the random part parameters to estimate regression coefficients, use the latter estimates to estimate the random part parameters again (and now better), then go on to estimate the regression coefficients again, and so on *ad libitum* – or, rather, until convergence of this iterative process. This loose description is close to the iterated generalized least squares (IGLS) method that is one of the algorithms for calculating the ML estimates; see Goldstein (2011).

There exist other methods for calculating the same estimates – Fisher scoring (Longford, 1993) and the EM algorithm (see Raudenbush and Bryk, 2002) – each with its own advantages. An extensive mathematical explanation is given in de Leeuw and Meijer (2008b).

Parameters can again be estimated with the ML or REML method; the REML method is preferable in the sense that in the case of small sample sizes (N) at the higher level it produces less biased estimates for the random part parameters and more reliable standard errors (Section 6.1). The ML method, however, is more convenient if one wishes to use deviance tests (see the next chapter). The IGLS algorithm produces ML estimates, whereas the so-called ‘restricted IGLS’ (RIGLS) algorithm yields the REML estimates. For the random slopes model it is also possible that estimates for the variance parameters τ_h^2 are exactly 0. The explanation is analogous to the explanation given for the intercept variances in Section 4.7.

The random group effects U_{hj} can again be ‘estimated’ by the empirical Bayes method, and the resulting ‘estimates’ are called posterior slopes (sometimes posterior means). This is analogous to the treatment of posterior means in Section 4.8.

Usually the estimation algorithms do not allow an unlimited number of random slopes to be included. Depending on the data set and the model specification, it is not uncommon for the algorithm to refuse to converge for more than two or three variables with random slopes. Sometimes the convergence can be improved by linearly transforming the variables with random slopes so that they have (approximately) zero means, or by transforming them to have (approximately) zero correlations.

For some data sets, the estimation method can produce estimated variance and covariance parameters that correspond to impossible covariance matrices for the random effects at level two. For example, τ_{01} sometimes is estimated to be larger than $\tau_0 \times \tau_1$, which would imply an intercept–slope correlation larger than 1. This is not an error of the estimation procedure, and it can be understood as follows. The estimation procedure is directed at the mean vector and covariance matrix of the vector of all observations. Some combinations of parameter values correspond to permissible structures of the latter covariance matrix that, nevertheless, cannot be formulated as a random effects model such as (5.3). Even if the estimated values for the τ_{hk} parameters do not combine into a positive definite matrix τ , the σ^2 parameter will still make the covariance matrix of the original observations (cf. equations (5.5) and (5.6)) positive definite. Therefore, such a strange result is in contradiction to the random effects formulation (5.3), but not to the more general formulation of a patterned covariance matrix for the observations.

Most computer programs give the standard errors of the variances of the random intercept and slopes; some give the standard errors of the estimated standard deviations $\hat{\tau}_h$ instead. This is related to the fact that the estimated variances may have very skewed distributions, especially if these distributions are close to the point 0. The estimated standard deviation (i.e., the square root of the estimated variance) will often have a more nearly normal distribution than the estimated variance itself. Approximate relations between the standard errors of estimated variances and estimated standard deviations are given in (6.2) in the next chapter.

5.5 Three or more levels

When the data have a three-level hierarchy, as discussed also in Section 4.9, slopes of level-one variables can be made random at level two and also at level three. In this case there will be at least two level-two and two level-three equations: one for the random intercept and one for the random slope. So, in the case of one explanatory variable, the model might be formulated as follows:

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk} x_{ijk} + R_{ijk}, \quad (\text{level-one model})$$

$$\beta_{0jk} = \delta_{00k} + U_{0jk}, \quad (\text{level-two model for intercept})$$

$$\beta_{1jk} = \delta_{10k} + U_{1jk}, \quad (\text{level-two model for slope})$$

$$\delta_{00k} = \gamma_{000} + V_{00k}, \quad (\text{level-three model for intercept})$$

$$\delta_{10k} = \gamma_{100} + V_{10k}. \quad (\text{level-three model for slope})$$

In the specification of such a model, for each level-one variable with random slope it has to be decided whether its slope must be random at level two, random at level three, or both.

Generally one should have either strong *a priori* knowledge or a good theory to formulate models as complex as this one or even more complex models (i.e., with more random slopes). Further, for each level-two variable it must be decided whether its slope is random at level three.

Example 5.5 *A three-level model with a random slope.*

We continue with the example of Section 4.9, where we illustrated the three-level model using a data set about a mathematics test administered to students in classes in schools. Now we include the available covariates (which are all centered around their grand means), and the regression coefficient for the mathematics pretest is allowed to be random at levels two and three. The results are shown in Table 5.5.

Table 5.5: Estimates for three-level model with random slopes.

Fixed effect	Coefficient	S.E.
γ_{000} = Intercept	8.41	0.16
γ_{100} = Coefficient of IQ	0.050	0.005
γ_{200} = Coefficient of pretest	0.146	0.011
γ_{300} = Coefficient of motivation	0.032	0.008
γ_{400} = Coefficient of father's education	0.039	0.015
γ_{500} = Coefficient of gender	0.221	0.106
Random part	Parameters	S.E.
<i>Level-three random part:</i>		
$\varphi_0^2 = \text{var}(V_{00k})$	0.971	0.254
$\varphi_2^2 = \text{var}(V_{20k})$	0.0024	0.0010
$\varphi_{02} = \text{cov}(V_{00k}, V_{20k})$	0.0381	0.0135
<i>Level-two random part:</i>		
$\tau_0^2 = \text{var}(U_{0jk})$	0.439	0.089
$\tau_2^2 = \text{var}(U_{2jk})$	0.0019	0.0009
$\tau_{02} = \text{cov}(U_{0jk}, U_{2jk})$	0.0398	0.0068
<i>Level-one variance:</i>		
$\sigma^2 = \text{var}(R_{ijk})$	5.978	0.145
Deviance	17,808.0	

The interpretation of the fixed part is straightforward as in conventional single-level regression models. The random part is more complicated. Since all predictor variables were grand mean centered, the intercept variances φ_0^2 (level three) and τ_0^2 (level two) have a clear meaning: they represent the amount of variation in mathematics achievement across schools and across classes within schools, respectively, for the average student whilst controlling for differences in IQ, mathematics ability, achievement motivation, educational level of the father, and gender. Comparing this table with Table 4.5 shows that much of the initial level-three and (especially) level-two variation has now been accounted for. Once there is control for initial differences, schools and classes within schools differ considerably less in the average mathematics achievement of their students at the end of grade 2.

Now we turn to the slope variance. The fixed slope coefficient for the mathematics pretest is estimated to be 0.146. The variance at level three for this slope is 0.0024, and at level two 0.0019. So the variability between schools of the effect of the pretest is somewhat larger than the variability of this effect between classes. At one end of the distribution there are a few percent of the schools that have an effect of the pretest that is only $0.146 - 2 \times \sqrt{0.0024} = 0.05$, whereas in the most selective schools this effect is $0.146 + 2 \times \sqrt{0.0024} = 0.252$.

There is also variation in this coefficient across classes within schools. In the population of classes, when we draw a random class in a random school, the total slope variance is $0.0024 + 0.0019$. Therefore, the classes with the 2.5% lowest and highest effect of pretest have regression coefficients $0.146 \pm 2\sqrt{0.0024 + 0.0019}$. Hence the gap between initial low and initial high achievers (4 standard deviations apart; this standard deviation for the pretest is 8.21) within the same class can become as large as $(0.146 + 2 \times \sqrt{0.0024 + 0.0019}) \times (4 \times 8.21) = 9$ points, or on the other hand it can remain as small as $(0.146 - 2 \times \sqrt{0.0024 + 0.0019}) \times (4 \times 8.21) = 1$ point in the least selective classes. Given the standard deviation of 3.4 for the dependent variable, a difference of 1 is very small, whereas 9 is quite a large difference.

5.6 Glommary

Hierarchical linear model. The main model for multilevel analysis, treated in this chapter.

Random slope. The random residual at level two in the hierarchical linear model indicating group-level deviations in the effect of an explanatory variable X on the dependent variable, symbolized in formula (5.3) by U_{lj} and in the more general formula (5.15) by U_{hj} . It is associated with the *slope variance*, the variance of U_{hj} .

To interpret the random slope variance, it is often helpful to consider the distribution of the group-dependent slopes. This distribution follows from

$$\beta_{hj} = \gamma_{h0} + U_{hj}$$

in a model such as (5.2) without cross-level interactions. This follows a normal distribution with mean γ_{h0} and variance τ_h^2 , the slope variance. If the random slope β_{hj} , a latent variable, is explained by group-level variables Z_1, \dots, Z_q , which leads to cross-level interactions in the model for Y_{ij} , the model for the random slope is given in (5.11),

$$\beta_{hj} = \gamma_{h0} + \gamma_{h1} z_{1j} + \dots + \gamma_{hq} z_{qj} + U_{hj}.$$

Heteroscedasticity. The dependent variable in a random slope model is heteroscedastic, that is, the residual standard deviation depends on the variables that have random slopes.

Cross-level interaction effect. An interaction between two variables defined at different levels, which can enter the model when a random slope is explained by a level-two variable.

Model specification. This received ample attention, including the following points:

1. Covariances between random intercepts and random slopes should (almost always) be estimated as free parameters, and not be equated *a priori* to zero.

2. Random slopes can often be explained by level-two variables, leading to models with cross-level interactions.
3. Decisions have to be made about whether or not to apply group centering to variables that have random slopes.
4. To avoid ecological fallacies and have the best possible theoretical interpretations, it is important also to consider interactions involving group means of level-one variables.

Parameter estimation. A brief intuitive overview of parameter estimation in the hierarchical linear model was given in Section 5.4.

Empirical Bayes estimation. The random slopes can be estimated by empirical Bayes methods just like the random intercepts. Such estimates are called *posterior slopes*.

Higher-level hierarchical linear models. The hierarchical linear model can also be defined for three- and higher-level data structures, and for the three-level case this was discussed in Section 5.5.

6

Testing and Model Specification

OVERVIEW OF THE CHAPTER

The first part of this chapter discusses hypothesis tests for parameters in the hierarchical linear model, with separate treatments of tests for the parameters in the fixed part and those in the random part. The second part is concerned with model specification issues.

It is assumed that the reader has a basic knowledge of statistical hypothesis testing: null and alternative hypotheses, type I and type II errors, significance level, and statistical power.

6.1 Tests for fixed parameters

Suppose we are working with a model represented as (5.15). The null hypothesis that a certain regression parameter is 0,

$$H_0 : \gamma_h = 0,$$

can be tested by a t -test. The statistical estimation leads to an estimate $\hat{\gamma}_h$ with associated standard error $S.E.(\hat{\gamma}_h)$. For a small number of groups $N \leq 50$ the standard errors based on REML estimation should be used (the difference between REML and ML estimation was explained in Sections 4.7 and 5.4; the need to use REML standard errors is argued by Manor and Zucker, 2004), while for a larger number of groups this is not important. Their ratio is a t -value:

$$T(\gamma_h) = \frac{\hat{\gamma}_h}{S.E.(\hat{\gamma}_h)} \tag{6.1}$$

One-sided as well as two-sided tests can be carried out on the basis of this test statistic.¹ Under the null hypothesis, $T(\gamma_h)$ has approximately a t distribution, but the number of degrees of freedom (df) is somewhat more complicated than in multiple linear regression,

¹This is one of the common principles for construction of a t -test. This type of test is called the *Wald test*, after the statistician Abraham Wald (1902–1950).

because of the presence of the two levels. The approximation by the t distribution is not exact even if the normality assumption for the random coefficients holds. Suppose first that we are testing the coefficient of a level-one variable. If the total number of level-one units is M and the total number of explanatory variables is r , then we can take $df = M - r - 1$. To test the coefficient of a level-two variable when there are N level-two units and q explanatory variables at level two, we take $df = N - q - 1$. To test the coefficient of the cross-level interaction between level-one variable X and level-two variable Z , when the model contains a total of q other level-two variables also interacting with this variable X , we also use $df = N - q - 1$.

If the number of degrees of freedom is large enough (say, larger than 40), the t distribution can be replaced by a standard normal distribution.

This rule for the degrees of freedom in the t -test is simple and has good properties. The literature contains proposals for various other rules. Manor and Zucker (2004) give a review and a simulation study. Their conclusion is, first, that these tests should use the standard errors obtained from REML estimation, not from ML estimation. Furthermore, they found that the so-called Satterthwaite approximation as well as the simple approximation mentioned here give a good control of the type I error rate, and most other methods give inflated type I error rates for fewer than 30 groups. Maas and Hox (2004) confirmed that for 30 or more groups, the Wald tests of fixed effects using REML standard errors have reliable type I error rates.

Example 6.1 Testing within- and between-group regressions.

Continuing the examples on students in schools of Chapters 4 and 5, we now wish to test whether between-group and within-group regressions of language test score on IQ are different from one another, when controlling for socio-economic status (SES). A model with a random slope for IQ is used. Two models are estimated and presented in Table 6.1. The first contains the raw (i.e., grand-mean-centered) IQ variable along with the group mean, the second contains the within-group deviation variable \tilde{IQ} , defined as

$$\tilde{IQ}_{ij} = IQ_{ij} - \bar{IQ}_{j\cdot}$$

also together with the group mean. The variable with the random slope is in both models the grand-mean-centered variable IQ. (It may be noted that although this variable is not listed among the variables with a fixed effect in Model 2, it is present implicitly because it is the sum of two included variables that do have a fixed effect.)

To test whether within- and between-group regression coefficients are different, the significance of the group mean \bar{IQ} is tested, while controlling for the effect of the original variable, IQ. The difference of within- and between-group regressions is discussed further in Section 4.6.

Table 6.1 shows that only the estimate for \bar{IQ} differs between the two models. This is in accordance with Section 4.6: if IQ is variable 1 and \bar{IQ} is variable 2, so that their regression coefficients are γ_{10} and γ_{20} , respectively, then the within-group regression coefficient is γ_{10} in Model 1 and γ_{20} in Model 2, while the between-group regression coefficient is $\gamma_{10} + \gamma_{20}$ in Model 1 and γ_{20} in Model 2. The models are equivalent representations of the data and differ only in the parametrization. The deviances (explained in Section 6.2) are exactly the same.

The within-group regression and between-group regressions are the same if $\gamma_{20} = 0$ in Model 1, that is, there is no effect of the group mean given that the model controls for the raw variable (i.e., the variable without group centering). The t -test statistic for testing $H_0 : \gamma_{20} = 0$ in Model 1 is equal to $0.647/0.264 = 2.45$. This is significant (two-sided $p < 0.02$). It may be concluded that within-group and between-group regressions are significantly different.

Table 6.1: Estimates for two models with different between- and within-group regressions.

	Model 1		Model 2	
Fixed effects	Coefficient	S.E.	Coefficient	S.E.
γ_{00} = Intercept	41.15	0.23	41.15	0.23
γ_{10} = Coefficient of IQ	2.265	0.065		
γ_{20} = Coefficient of \bar{IQ}			2.265	0.065
γ_{30} = Coefficient of SES	0.161	0.011	0.161	0.011
γ_{01} = Coefficient of \bar{IQ}	0.647	0.264	2.912	0.262
Random part	Parameter	S.E.	Parameter	S.E.
<i>Level-two parameters:</i>				
$\tau_0^2 = \text{var}(U_{0j})$	9.08	1.12	9.08	1.12
$\tau_1^2 = \text{var}(U_{1j})$	0.197	0.074	0.197	0.074
$\tau_{01} = \text{cov}(U_{0j}, U_{1j})$	-0.815	0.214	-0.815	0.214
<i>Level-one variance:</i>				
$\sigma^2 = \text{var}(R_{ij})$	37.42	0.91	37.42	0.91
Deviance	24,661.3		24,661.3	

The results for Model 2 can be used to test whether the within-group or between-group regressions are 0. The t -statistic for testing the within-group regression is $2.265/0.065 = 34.9$, the statistic for testing the between-group regression is $2.912/0.262 = 11.1$. Both are extremely significant. In conclusion, there are positive within-group as well as between-group regressions, and these are different from one another.

6.1.1 Multiparameter tests for fixed effects

Sometimes we wish to test several regression parameters simultaneously. For example, consider testing the effect of a categorical variable with more than three categories. The effect of such a variable can be represented in the fixed part of the hierarchical linear model by $c - 1$ dummy variables, where c is the number of categories, and this effect is nil if and only if all the corresponding $c - 1$ regression coefficients are 0. Two types of test are much used for this purpose: the multivariate Wald test and the likelihood ratio test, also known as the *deviance test*. The latter test is explained in the next section.

For the multivariate Wald test, we need not only the standard errors of the estimates but also the covariances among them. Suppose that we consider a certain vector γ of q regression parameters, for which we wish to test the null hypothesis

$$H_0 : \gamma = 0.$$

The statistical estimation leads to an estimate $\hat{\gamma}$ and an associated estimated covariance matrix $\hat{\Sigma}_{\gamma}$. Here again the covariance matrix produced by REML estimation, not by ML estimation, needs to be used, unless the number of groups is large enough ($N \geq 50$). From

the estimate and its covariance matrix, we can let a computer program calculate the test statistic represented in matrix form by

$$\hat{\gamma}' \hat{\Sigma}_{\gamma}^{-1} \hat{\gamma}.$$

Under the null hypothesis, the distribution of this statistic divided by q can be approximated by an F distribution, with q degrees of freedom in the numerator, while the degrees of freedom in the denominator are determined as for the t -test explained above. Here, for a large number of degrees of freedom (say, more than 40), the F distribution is approximated by a chi-squared distribution with q degrees of freedom.

The way of obtaining tests presented in this section is not applicable to tests of whether parameters (variances or covariances) in the random part of the model are 0. The reason is that, if a population variance parameter is 0, its estimate divided by the estimated standard error does not approximately have a t distribution. Tests for such hypotheses are discussed in the next section.

6.2 Deviance tests

The deviance test, or likelihood ratio test, is a quite general principle for statistical testing. In applications of the hierarchical linear model this test is mainly used for multiparameter tests and for tests about the random part of the model. The general principle is as follows.

When parameters of a statistical model are estimated by the maximum likelihood method, the estimation also provides the likelihood, which can be transformed into the *deviance* defined as minus twice the natural logarithm of the likelihood. This deviance can be regarded as a measure of lack of fit between model and data, but (in most statistical models) one cannot interpret the values of the deviance directly, but only differences in deviance values for several models fitted to the same data set.

Suppose that two models are fitted to one data set, model M_0 with m_0 parameters and a larger model M_1 with m_1 parameters. Thus M_1 can be regarded as an extension of M_0 , with $m_1 - m_0$ parameters added. Suppose that M_0 is tested as the null hypothesis and M_1 is the alternative hypothesis. Indicating the deviances by D_0 and D_1 , respectively, their difference $D_0 - D_1$ can be used as a test statistic having a chi-squared distribution with $m_1 - m_0$ degrees of freedom. This type of test can be applied to parameters of the fixed as well as of the random part.

The deviance produced by the residual maximum likelihood method can be used in deviance tests only if the two models compared (M_0 and M_1) have the same fixed parts and differ only in their random parts. Otherwise the ML estimation method must be used.

Example 6.2 Test of random intercept.

In Example 4.2, the random intercept model yields a deviance of $D_1 = 24,912.2$, while the OLS regression model has a deviance of $D_0 = 25,351.0$. There is $m_1 - m_0 = 1$ parameter added, the random intercept variance. The deviance difference is 438.8, immensely significant in a chi-squared distribution with $df = 1$. This implies that, even when controlling for the effect of IQ, the differences between groups are strongly significant.

In Section 6.2.1 we shall see, however, that the test of the random intercept can be improved.

For example, suppose one is testing the fixed effect of a categorical explanatory variable with c categories. This categorical variable can be represented by $c - 1$ dummy variables.

Model M_0 is the hierarchical linear model with the effects of the other variables in the fixed part and with the given random part. Model M_1 also includes all these effects; in addition, the $c - 1$ regression coefficients of the dummy variables have been added. Hence the difference in the number of parameters is $m_1 - m_0 = c - 1$. This implies that the deviance difference $D_0 - D_1$ can be tested against a chi-squared distribution with $df = c - 1$. This test is an alternative for the multiparameter Wald test treated in the preceding section. These tests will be very close to each other for intermediate and large sample sizes. For a small number of groups ($N < 40$) the precision of type I error rates can be improved by the so-called Bartlett correction (see Zucker et al., 2000; Manor and Zucker, 2004).

Example 6.3 Effect of a categorical variable.

In the data set used in Chapters 4 and 5, schools differ according to their denomination: public, Catholic, Protestant, nondenominational private, and a remainder category. To represent these five categories, four dummy variables are used, contrasting the last four categories with the first. This means that all dummy variables are 0 for public schools, the first is 1 for Catholic schools, etc. When the fixed effects of these $c - 1 = 4$ variables are added to the model presented in Table 5.2, which in this example has the role of M_0 with deviance $D_0 = 24,856.8$, the deviance decreases to $D_1 = 24,840.3$. The chi-squared value is $D_0 - D_1 = 16.5$ with $df = 4$, $p < 0.005$. It can be concluded that, while controlling for IQ, the group average of IQ, and the interaction of these two variables (see the specification of model M_0 in Table 5.2), there are differences between the five types of school.

The estimated fixed effects (with standard errors) of the dummy variables are 1.65 (0.55) for the Catholic and -0.51 (0.59) for the Protestant schools, and 0.60 (1.05) and 0.23 (0.94) for the other two categories. These effects are relative to the public schools. Calculating the t -ratios shows that the Catholic schools show higher achievement, controlling for these three IQ variables, than the public schools, while the other three categories do not differ significantly from the public schools.

6.2.1 More powerful tests for variance parameters

Variances are by definition nonnegative. When testing the null hypothesis that the variance of the random intercept or of a random slope is zero, the alternative hypothesis is therefore one-sided. This observation can be used to arrive at a sharpened version of the deviance test for variance parameters. This principle was derived by Miller (1977), Self and Liang (1987), and Stram and Lee (1994). Molenberghs and Verbeke (2007) give an overview of this and other tests for parameters in the random part. The good properties of this procedure were confirmed in a simulation study by LaHuis and Ferguson (2009).

First, consider the case where a random intercept is tested. The null model M_0 is the model without a random part at level two, that is, all observations Y_{ij} are independent, conditional on the values of the explanatory variables. This is an ordinary linear regression model. The alternative model M_1 is the random intercept model with the same explanatory variables. There is $m_1 - m_0 = 1$ additional parameter, the random intercept variance τ_0^2 . For the observed deviances D_0 of model M_0 (this model can be estimated by ordinary least squares) and D_1 of the random intercept model, the difference $D_0 - D_1$ is calculated. If $D_0 - D_1 = 0$, the random intercept variance is definitely not significant (it is estimated as being 0 ...). If $D_0 - D_1 > 0$, the tail probability of the difference $D_0 - D_1$ is looked up in a table of the chi-squared distribution with $df = 1$. The p -value for testing the significance of the random intercept variance is half this tail value.

Second, consider the case of testing a random slope. Specifically, suppose that model (5.15) holds, and the null hypothesis is that the last random slope variance is zero: $\tau_p^2 = 0$. Under this null hypothesis, the p covariances, τ_{hp} for $h = 0, \dots, p - 1$, are also 0. The alternative hypothesis is the model defined by (5.15), and has $m_1 - m_0 = p + 1$ parameters more than the null model (one variance and p covariances). For example, if there are no other random slopes in the model ($p = 1$), $m_1 - m_0 = 2$. A similar procedure is followed as for testing the random intercept, but now it is more complicated.² Both models are estimated, yielding the deviance difference $D_0 - D_1$. If $D_0 - D_1 = 0$, the random slope variance is not significant. If $D_0 - D_1 > 0$, the tail probability of the difference $D_0 - D_1$ can be tested against a so-called mixture distribution of the chi-squared distribution with $df = p + 1$ and the chi-squared distribution with $df = p$, with mixture proportions $\frac{1}{2}$. This is also called the chi-bar distribution. It is not a standard distribution. The p -value is just the average of the p -value from the χ_p^2 distribution and that from the χ_{p+1}^2 distribution. This can be easily calculated by any program that can calculate p -values (or cumulative distribution functions) for chi-squared distributions.

Table 6.2 gives critical values for testing a random slope in a model that has a random intercept and no, one, or two additional random slopes. These critical values are in between the critical values of the χ_p^2 and χ_{p+1}^2 distributions, where p is the total number of random slopes in the model including the tested one but excluding the random intercept. Therefore, a conservative test is obtained by testing against the χ_{p+1}^2 distribution.

Table 6.2: Critical values for 50–50 mixture of χ_p^2 and χ_{p+1}^2 distribution.

	α			
p	0.10	0.05	0.01	0.001
1	3.81	5.14	8.27	12.81
2	—	5.53	7.05	10.50
3	7.09	8.76	12.48	17.61

Example 6.4 Test of random slope.

When comparing Tables 4.4 and 5.1, it can be concluded that $m_1 - m_0 = 2$ parameters are added and the deviance diminishes by $D_0 - D_1 = 15,227.5 - 15,213.5 = 14.0$. Testing the value of 14.0 in Table 6.2 for $p = 1$ yields $p < 0.001$. Thus, the significance probability of the random slope for IQ in the model of Table 5.1 is $p < 0.001$.

As another example, suppose that one wishes to test the significance of the random slope for IQ in the model of Table 5.4. Then the model must be fitted in which this effect is omitted and all other effects remain. Thus, the omitted parameters are τ_1^2 and τ_{01} so that $df = 2$. Fitting this reduced model leads to a deviance of $D_0 = 24,655.1$, which is 28.3 more than the deviance in Table 5.4. Table 6.2 for $p = 1$ shows that again $p < 0.001$. Thus there is strong support for the random slope.

²In the first edition, we erroneously stated that exactly the same procedure could be followed.

6.3 Other tests for parameters in the random part

Deviance tests are very convenient for testing parameters in the random part, but other tests for random intercepts and slopes also exist. Some of these are reviewed by Molenberghs and Verbeke (2007). We mentioned in Section 3.3.2 the ANOVA F -test for the intraclass correlation. This is effectively a test for randomness of the intercept. If it is desired to test the random intercept while controlling for explanatory variables, one may use the F -test from the ANCOVA model, using the explanatory variables as covariates.

Raudenbush and Bryk (2002) present chi-squared tests for random intercepts and slopes, that is, for variance parameters in the random part. These are based on calculating OLS estimates for the values of the random effects within each group and testing these values for equality. For the random intercept, these are the large-sample chi-squared approximations to the F -tests in the ANOVA or ANCOVA model mentioned in Section 3.3.2.

Another test for variance parameters in the random part was proposed by Berkhof and Snijders (2001) and Verbeke and Molenberghs (2003), and is explained in Section 10.2.2.

6.3.1 Confidence intervals for parameters in the random part

When constructing confidence intervals for variance parameters it is tempting to use the symmetric confidence interval, which for a confidence level of 95% is defined as the estimate plus or minus 1.96 times the standard error. This confidence interval, however, is based on the assumption that the parameter estimate is nearly normally distributed, which is doubtful for estimated variances; for example, because these are necessarily nonnegative. These confidence intervals are applicable here only when the standard error is quite small compared to the estimate, so that they do not come near the value of 0. Generally, they are more safely applied to the standard deviation (the square root of the variance). A symmetric confidence interval for the standard deviation then can be transformed back to a nonsymmetric confidence interval for the variance, if this is desired. Sometimes also the logarithm of the variance can be used. (Note that the properties of the logarithm imply that $\ln(\tau^2) = 2 \ln(\tau)$.)

The standard errors of these transformations are related through the approximations³

$$\text{S.E.}(\ln(\hat{\tau}^2)) \approx \frac{\text{S.E.}(\hat{\tau}^2)}{\hat{\tau}^2}, \quad \text{S.E.}(\hat{\tau}) \approx \frac{\text{S.E.}(\hat{\tau}^2)}{2\hat{\tau}}, \quad (6.2)$$

where \ln denotes the natural logarithm. If this value is less than 0.1 the distribution of $\hat{\tau}$ will usually be close to a normal distribution; if it is between 0.1 and 0.3 then the symmetric confidence interval still will give a reasonable rough approximation; if it is larger than 0.3 (which means that the number of higher-level units is probably rather low, or the higher-level variance is quite small) the normal approximation itself will break down and the symmetric confidence interval should not be used.

A better way to construct the confidence interval is based on the so-called *profile likelihood*.⁴ This confidence interval consists of those values of the parameter for which

³These approximations are based on linear approximations to the logarithmic and square root functions (the so-called delta method), and are invalid if these standard errors are too large.

⁴The profile likelihood function for a given parameter is the value of the likelihood function for this parameter, maximized over all other parameters.

twice the logarithm of the profile likelihood is not smaller than twice the logarithm of the maximized likelihood in the ML estimate, minus a value c ; this c is the critical value in a chi-squared distribution on 1 degree of freedom ($c = 3.84$ for a confidence level of 95%). This confidence interval will be implemented in R package lme4⁵ (see Bates, 2010). Another option is to follow a Bayesian approach (Section 12.1) and construct a confidence interval based on the posterior distribution of τ^2 or τ without relying on normal approximations.

Example 6.5 Confidence interval for intercept and slope variances

In the random intercept model of Example 4.3, the parameter estimates for the two variance components are $\hat{\tau}_0^2 = 8.68$ (S.E. = 1.10) for the level-two variance and $\hat{\sigma}^2 = 40.43$ (S.E. = 0.96) for the level-one variance. The 95% confidence intervals based on the profile likelihood, obtained from R package lme4a, are as follows. The nature of the likelihood-based method implies that the endpoints of the confidence interval for the variances are the squares of those for the standard deviations: for the standard deviations, the interval is

$$2.60 \leq \tau_0 \leq 3.34, \quad 6.21 \leq \sigma \leq 6.51;$$

and for the variances,

$$6.77 \leq \tau_0^2 \leq 11.16, \quad 38.60 \leq \sigma^2 \leq 42.37.$$

The symmetric 95% confidence intervals based on the standard errors are defined as the parameter estimates plus or minus 1.96 times the standard errors. These confidence intervals are less reliable. In this example we use for the variances the standard errors in Table 4.4, and for the standard deviations the standard errors from equation (6.2). The results for the standard deviations are

$$2.58 \leq \tau_0 \leq 3.31, \quad 6.21 \leq \sigma \leq 6.51;$$

and for the variances,

$$6.53 \leq \tau_0^2 \leq 10.83, \quad 38.55 \leq \sigma^2 \leq 42.31.$$

It can be concluded that in this case, where the ratios of the variance estimates to their standard errors are reasonably large for both parameters, the symmetric confidence intervals are very good approximations when based on the standard deviations, especially for σ , and less good for the variances but for σ^2 still totally acceptable.

To construct a confidence interval for the intercept variance τ_0^2 using only the information in Table 4.4, if one does not have access to the profile likelihood method, the best way (as mentioned above) is to construct a symmetric confidence interval for the intercept standard deviation τ_0 and transform this to a confidence interval for the variance. This operates as follows (we work to three decimal places so as to have precision to two decimal places in the result):

- * From $\hat{\tau}_0^2 = 8.680$, calculate $\hat{\tau}_0 = 2.946$.
- * From $S.E.(\hat{\tau}_0^2) = 1.096$ and formula (6.2), calculate $S.E.(\hat{\tau}_0) = 1.096/(2 \times 2.946) = 0.1860$.
- * The confidence interval for τ_0 now extends from $2.946 - 1.960 \times 0.1860 = 2.581$ to $2.946 + 1.960 \times 0.1860 = 3.311$.
- * Hence the confidence interval for τ_0^2 extends from $2.581^2 = 6.66$ to $3.311^2 = 10.96$. This is still not as good as the interval based on the profile likelihood, but considerably better than the symmetric interval directly based on the estimated intercept variance.

⁵At the time of writing, it is implemented in the beta version lme4a, available from R-Forge.

6.4 Model specification

Model specification is the choice of a satisfactory model. For the hierarchical linear model this amounts to the selection of relevant explanatory variables (and interactions) in the fixed part, and relevant random slopes (with their covariance pattern) in the random part. Model specification is one of the most difficult parts of statistical inference, because there are two steering wheels: substantive (subject-matter related) and statistical considerations. These steering wheels must be handled jointly. The purpose of model specification is to arrive at a model that describes the observed data to a satisfactory extent but without unnecessary complications. A parallel purpose is to obtain a model that is of substantive interest without wringing from the data drops that are really based on chance but interpreted as substance. In linear regression analysis, model specification is already complicated – as is discussed in many textbooks on regression (e.g., Ryan, 1997) – and in multilevel analysis the number of complications is multiplied because of the complicated nature of the random part.

The complicated nature of the hierarchical linear model, combined with the two steering wheels for model specification, implies that there are no clear and fixed rules to follow. Model specification is a process guided by the following principles.

1. Considerations relating to the subject matter. These follow from field knowledge, existing theory, detailed problem formulation, and common sense.
2. The distinction between effects that are indicated *a priori* as effects to be tested, that is, effects on which the research is focused, and effects that are necessary to obtain a good model fit. Often the effects tested are a subset of the fixed effects, and the random part is to be fitted adequately but of secondary interest. When there is no strong prior knowledge about which variables to include in the random part, one may follow a data-driven approach to select the variables for the random part.
3. A preference for ‘hierarchical’ models in the general sense (not the ‘hierarchical linear model’ sense) that if a model contains an interaction effect, then the corresponding main effects should usually also be included (even if these are not significant); and if a variable has a random slope, its fixed effect should normally also be included in the model. The reason is that omitting such effects may lead to erroneous interpretations.
4. Doing justice to the multilevel nature of the problem. This is done as follows:
 - (a) When a given level-one variable is present, one should be aware that the within-group regression coefficient may differ from the between-group regression coefficient, as described in Section 4.6. This can be investigated by calculating a new variable, defined as the group mean of the level-one variable, and testing the effect of the new variable.
 - (b) When there is an important random intercept variance, there are important unexplained differences between group means. One may look for level-two variables (original level-two variables as well as aggregates of level-one variables) that explain part of these between-group differences.
 - (c) When there is an important random slope variance of some level-one variable (say, X_1), there are important unexplained differences between the within-group

effects of X_1 on Y . Here also one may look for level-two variables that explain part of these differences. This leads to cross-level interactions as explained in Section 5.2. Recall from this section, however, that cross-level interactions can also be expected from theoretical considerations, even if no significant random slope variance is found.

5. Awareness of the necessity of including certain covariances of random effects. Including such covariances means that they are free parameters in the model, not constrained to 0 but estimated from the data.

In Section 5.1.2, attention was given to the necessity to include in the model all covariances τ_{0h} between random slopes and random intercept. Another case in point arises when a categorical variable with $c \geq 3$ categories has a random effect. This is implemented by giving random slopes to the $c - 1$ dummy variables that are used to represent the categorical variable. The covariances between these random slopes should then also be included in the model.

Formulated generally, suppose that two variables X_h and $X_{h'}$ have random effects, and that the meaning of these variables is such that they could be replaced by two linear combinations, $aX_h + a'X_{h'}$ and $bX_h + b'X_{h'}$. (For the random intercept and random slope discussed in Section 5.1.2, the relevant type of linear combination would correspond to a change of origin of the variable with the random slope.) Then the covariance $\tau_{hh'}$ between the two random slopes should be included in the model.

6. Reluctance to include nonsignificant effects in the model – one could also say, a reluctance to overfit. Each of points 1–5 above, however, could override this reluctance.

An obvious example of this overriding is the case where one wishes to test for the effect of X_2 , while controlling for the effect of X_1 . The purpose of the analysis is a subject-matter consideration, and even if the effect of X_1 is nonsignificant, one still should include this effect in the model.

7. The desire to obtain a good fit, and include all effects in the model that contribute to a good fit. In practice, this leads to the inclusion of all significant effects unless the data set is so large that certain effects, although significant, are deemed unimportant nevertheless.
8. Awareness of the following two basic statistical facts:

- (a) Every test of a given statistical parameter controls for all other effects in the model used as a null hypothesis (M_0 in Section 6.2). Since the latter set of effects has an influence on the interpretation as well as on the statistical power, test results may depend on the set of other effects included in the model.
- (b) We are constantly making type I and type II errors. Especially the latter, since statistical power often is rather low. This implies that an effect being nonsignificant does not mean that the effect is absent in the population. It also implies that a significant effect may be so by chance (but the probability of this is no larger than the level of significance – most often set at 0.05).

Multilevel research often is based on data with a limited number of groups. Since power for detecting effects of level-two variables depends strongly on the number of groups in the data, warnings about low power are especially important for level-two variables.

9. Providing tested fixed effects with an appropriate error term in the model (whether or not it is significant). For level-two variables, this is the random intercept term (the residual term in (5.7)). For cross-level interactions, it is the random slope of the level-one variable involved in the interaction (the residual term in (5.8)). For level-one variables, it is the regular level-one residual that one would not dream of omitting. This guideline is supported by Berkhof and Kampen (2004).

These considerations are nice – but how should one proceed in practice? To get an insight into the data, it is usually advisable to start with a descriptive analysis of the variables: an investigation of their means, standard deviations, correlations, and distributional forms. It is also helpful to make a preliminary ('quick and dirty') analysis with a simpler method such as OLS regression.

When starting with the multilevel analysis as such, in most situations (longitudinal data may provide an exception), it is advisable to start with fitting the empty model (4.6). This gives the raw within-group and between-group variances, from which the estimated intraclass correlation can be calculated. These parameters are useful as a general description and a starting point for further model fitting. The process of further model specification will include *forward steps* – select additional effects (fixed or random), test their significance, and decide whether or not to include them in the model – and *backward steps* – exclude effects from the model because they are not important from a statistical or substantive point of view. We mention two possible approaches in the following subsections.

6.4.1 Working upward from level one

In the spirit of Section 5.2, one may start with the construction of a model for level one, that is, first explaining within-group variability, and then explaining between-group variability. This two-phase approach is advocated by Raudenbush and Bryk (2002) and is followed in the HLM program (see Chapter 18).

Modeling within-group variability

Subject-matter knowledge and availability of data lead to a number of level-one variables X_1, \dots, X_p which are deemed important, or hypothetically important, to predict or explain the value of Y . These variables lead to equation (5.10) as a starting point:

$$Y_{ij} = \beta_{0j} + \beta_{1j} x_{1ij} + \dots + \beta_{pj} x_{prij} + R_{ij}.$$

This equation represents the within-group effects of X_h on Y . The between-group variability is first modeled as random variability. This is represented by splitting the group-dependent regression coefficients β_{hj} into a mean coefficient γ_{h0} and a group-dependent deviation U_{hj} :

$$\beta_{hj} = \gamma_{h0} + U_{hj}.$$

Substitution yields

$$Y_{ij} = \gamma_{00} + \sum_{h=1}^p \gamma_{h0} x_{hij} + U_{0j} + \sum_{h=1}^p U_{hj} x_{hij} + R_{ij}. \quad (6.3)$$

This model has a number of level-one variables with fixed and random effects, but it will usually not be necessary to include all random effects.

For the precise specification of the level-one model, the following steps are useful.

1. Select in any case the variables on which the research is focused. In addition, select relevant available level-one variables on the basis of subject-matter knowledge. Also include plausible interactions between level-one variables.
2. Select, among these variables, those for which, on the basis of subject-matter knowledge, a group-dependent effect (random slope!) is plausible. If unsure, one could select the variables that are expected to have the strongest fixed effects.
3. Estimate the model with the fixed effects of step 1 and the random effects of step 2.
4. Test the significance of the random slopes, and exclude the nonsignificant slopes from the model.
5. Test the significance of the regression coefficients, and exclude the nonsignificant coefficients from the model. This may also be the moment to consider the inclusion of interaction effects between level-one variables.
6. As a check, one could test whether the variables for which a group-dependent effect (i.e., a random slope) was not thought plausible in step 2 do indeed have a nonsignificant random slope. (Keep in mind that including a random slope normally implies inclusion of the fixed effect!) Be reluctant to include random slopes for interactions; these are often hard to interpret.
7. Use the methods of Chapter 10 for checking the assumptions (such as homoscedasticity and normal distributions of the residuals at all levels).

With respect to the random slopes, one may be restricted by the fact that data usually contain less information about random effects than about fixed effects. Including many random slopes can therefore lead to long iteration processes of the estimation algorithm. The algorithm may even fail to converge. For this reason it may be necessary to specify only a small number of random slopes.

After this process, one has arrived at a model with a number of level-one variables, some of which have a random effect in addition to their fixed effect. It is possible that the random intercept is the only remaining random effect. This model is an interesting intermediate product, as it indicates the within-group regressions and their variability.

Modeling between-group variability

The next step is to try and explain these random effects by level-two variables. The random intercept variance can be explained by level-two variables, the random slopes by interactions of level-one with level-two variables, as was discussed in Section 5.2. It should be kept in mind that aggregates of level-one variables can be important level-two variables.

For deciding which main effects of level-two variables and which cross-level interactions to include, it is again advisable first to select those effects that are plausible on the basis of substantive knowledge, then to test these and include or omit them depending on their importance (statistical and substantive), and finally to check whether other (less plausible) effects also are significant.

This procedure has a built-in filter for cross-level interactions: an interaction between level-one variable X and level-two variable Z is considered only if X has a significant random slope. However, this ‘filter’ should not be employed as a strict rule. If there are theoretical reasons to consider the $X \times Z$ interaction, this interaction can be tested even if X does not have a significant random slope. The background to this is the fact that if there is $X \times Z$ interaction, the test for this interaction has a higher power to detect this than the test for a random slope.

It is possible that, if one carries out both tests, the test of the random slope is non-significant while the test of the $X \times Z$ interaction is significant. This implies that either a type I error has been made by the test on $X \times Z$ interaction (this is the case if there is no interaction), or a type II error has been made by the test of the random slope (this is the case if there is interaction). Assuming that the significance level is 0.05 and one focuses on the test of this interaction effect, the probability of a type I error is less than 0.05 whereas the probability of a type II error can be quite high, especially since the test of the random slope does not always have high power for testing the specific alternative hypothesis of an $X \times Z$ interaction effect. Therefore, provided that the $X \times Z$ interaction effect was hypothesized before looking at the data, the significant result of the test of this effect is what counts, and not the lack of significance for the random slope.

By the same reasoning, if there is theoretical justification for testing the effect of a level-two variable it is not necessary to require the precondition that there is positive and significant random intercept variance.

6.4.2 Joint consideration of level-one and level-two variables

The procedure of first building a level-one model and then extending it with level-two variables is neat but not always the most efficient or most relevant. If there are level-two variables or cross-level interactions that are known to be important, why not include them in the model right from the start? For example, it could be expected that a certain level-one variable has a within-group regression differing from its between-group regression. In such a case, one may wish to include the group mean of this variable at the outset.

In this approach, the same steps are followed as in the preceding section, but without the distinction between level-one and level-two variables. This leads to the following steps:

1. Select relevant available level-one and level-two variables on the basis of subject-matter knowledge. Also include plausible interactions. Do not forget group means of level-one variables to account for the possibility of difference between within-group and between-group regressions. Also do not forget cross-level interactions.
2. Select among the level-one variables those for which, on the basis of subject-matter knowledge, a group-dependent effect (random slope!) is plausible. (A possibility would be, again, to select those variables that are expected to have the strongest fixed effects.)

3. Estimate the model with the fixed effects of step 1 and the random effects of step 2.
4. Test the significance of the random slope variance, and exclude from the model the random slopes of those variables for which this variance is not significant.
5. Test the significance of the regression coefficients, and exclude the nonsignificant coefficients from the model. This can also be a moment to consider the inclusion of more interaction effects.
6. Check whether other effects, thought less plausible at the start of model building, are indeed not significant. If they are significant, include them in the model.
7. Use the methods of Chapter 10 for checking the assumptions (such as homoscedasticity and normal distributions of the residuals at all levels).

In an extreme instance of step 1, one may wish to include all available variables and a large number of interactions in the fixed part. Similarly, one might wish to give all level-one variables random effects in step 2. Whether this is possible in practice will depend, *inter alia*, on the number of level-one variables. Such an implementation of these steps leads to a backward model-fitting process, where one starts with a large model and reduces it by stepwise excluding nonsignificant effects. The advantage is that masking effects (where a variable is excluded early in the model building process because of nonsignificance, whereas it would have reached significance if one had controlled for another variable) do not occur. The disadvantage is that it may be a very time-consuming procedure.

6.4.3 Concluding remarks on model specification

This section has suggested a general approach to specification of multilevel models rather than laying out a step-by-step procedure. This is in accordance with our view of model specification as a process with two steering wheels and without foolproof procedures. This implies that, given one data set, a (team of) researcher(s) may come up with more than one model, each in itself an apparently satisfactory result of a model specification process. In our view, this reflects the basic indeterminacy that is inherent in model fitting on the basis of empirical data. It is quite possible for several different models to correspond to a given data set, and that there are no compelling arguments to choose between them. In such cases, it is better to accept this indeterminacy and leave it to be resolved by further research than to make an unwarranted choice between the different models.

This treatment of model specification may seem rather inductive, or data-driven. If one is in the fortunate situation of having *a priori* hypotheses to test (usually about regression coefficients), it is useful to distinguish between, on the one hand, the parameters on which hypothesis tests are focused, and, on the other hand, the parts of the model that are required to make it fit the data well. The latter part is important to get right in order to obtain valid tests of the hypotheses. These two parts can be treated distinctly: the first part must evidently be included in the model, and for the second part an inductive approach is adequate.

Another aspect of model specification is the checking of assumptions. Independence assumptions should be checked in the course of specifying the random part of the model. Distributional assumptions – specifically, the assumption of normal distributions for the

various random effects – should be checked by residual analysis. Checks of assumptions are treated in Chapter 10.

6.5 Glommary

***t*-test.** The easiest test for a fixed coefficient is the so-called Wald test. Its test statistic is the *t*-ratio, that is, the ratio of estimate to standard error. This ratio is tested against a *t* distribution. For small N the number of degrees of freedom is important and rules were given for this purpose.

***F*-test.** Multiparameter tests of fixed coefficients according to the principle of the Wald test lead to *F*-tests, which for large N can be tested against a chi-squared distribution.

Deviance test. A general procedure for hypothesis testing is the deviance test or likelihood ratio test. This can be easily used for multiparameter tests, for example, tests of a categorical variable in the fixed part of the model.

Random intercept and random slope variances can also be tested by deviance tests. Then some power can be gained by using the fact that variances cannot be negative, which leads to a modified null distribution.

ML and REML estimation. For Wald tests of parameters in the fixed part when the number of higher-level units is rather small ($N \leq 50$), REML standard errors must be used. For deviance tests of random part parameters, ML estimates must be used.

Confidence intervals for random part parameters. These are tricky, because the estimated variances mostly do not have nearly normal distribution so that the usual symmetric confidence interval is inapplicable. Accordingly, the interpretation of standard errors of parameters in the random part is dubious unless the standard error is quite small compared to the parameter estimate. The symmetric 95% confidence interval, defined as parameter estimate ± 1.96 times the standard error, is more precise when applied to the standard deviation (τ) than to the variance (τ^2); this approach requires that the resulting interval is far enough away from 0. The profile likelihood method provides better confidence intervals, without assuming normality for the parameter estimates. Another possibility is to follow a Bayesian approach (Section 12.1) and construct a posterior distribution.

Model specification. This depends on a combination of subject-matter and statistical considerations, and it is impossible to give simple rules. A number of considerations and possible approaches were discussed, and more will follow in Chapter 10.

7

How Much Does the Model Explain?

OVERVIEW OF THE CHAPTER

This chapter starts by noting that, counter to what might intuitively be expected, estimated variance parameters may go up when explanatory variables are added to the model. This leads to issues in the definition of the concept of explained variance, commonly referred to as R^2 . A definition of R^2 , defined at level one, is given which does not have this problem in its population values, although it still may sometimes have the problem in values calculated from data.

Next an exposition is given of how different components of the hierarchical linear model contribute to the total observed variance of the dependent variable. This is a rather theoretical section that may be skipped by the reader.

7.1 Explained variance

The concept of ‘explained variance’ is well known in multiple regression analysis: it gives an answer to the question how much of the variability of the dependent variable is accounted for by the linear regression on the explanatory variables. The usual measure for the *explained proportion of variance* is the squared multiple correlation coefficient, R^2 . For the hierarchical linear model, however, the concept of ‘explained proportion of variance’ is somewhat problematic. In this section, we follow the approach of Snijders and Bosker (1994) to explain the difficulties and give a suitable multilevel version of R^2 .

One way to approach this concept is to transfer its customary treatment, well-known from multiple linear regression, straightforwardly to the hierarchical random effects model: treat proportional reductions in the estimated variance components, σ^2 and τ_0^2 in the random-intercept model for two levels, as analogs of R^2 -values. Since there are several variance components in the hierarchical linear model, this approach leads to several R^2 -values, one for each variance component. However, this definition of R^2 now and then leads to unpleasant surprises: it sometimes happens that adding explanatory variables *increases* rather than decreases some of the variance components. Even negative values of R^2 are possible. Negative values of R^2 are clearly undesirable and are not in accordance with its intuitive interpretation.

In the discussion of R^2 -type measures, it should be kept in mind that these measures depend on the distribution of the explanatory variables. This implies that these variables, denoted in this section by X , are supposed to be drawn at random from the population at level one and the population at level two, and not determined by the experimental design or the researcher's whim. In order to stress the random nature of the X -variable, the values of X are denoted X_{ij} , instead of by x_{ij} as in earlier chapters.

7.1.1 Negative values of R^2 ?

As an example, we consider data from a study by Vermeulen and Bosker (1992) on the effects of part-time teaching in primary schools. The dependent variable Y is an arithmetic test score; the sample consists of 718 grade 3 pupils in 42 schools. An intelligence test score X is used as predictor variable. Group sizes range from 1 to 33 with an average of 20. In the following, it sometimes is desirable to present an example for balanced data (i.e., with equal group sizes). The balanced data presented below are the data artificially restricted to 33 schools with 10 pupils in each school, by deleting schools with fewer than 10 pupils from the sample and randomly sampling 10 pupils from each school from the remaining schools. For demonstration purposes, three models were fitted: the empty Model A; Model B, with a group mean, \bar{X}_j , as predictor variable; and Model C, with a within-group deviation score ($X_{ij} - \bar{X}_j$) as predictor variable. Table 7.1 presents the results of the analyses both for the balanced and for the entire data set. The residual variance at level one is denoted σ^2 ; the residual variance at level two is denoted τ_0^2 .

Table 7.1: Estimated residual variance parameters $\hat{\sigma}^2$ and $\hat{\tau}_0^2$ for models with within-group and between-group predictor variables.

	$\hat{\sigma}^2$	$\hat{\tau}_0^2$
I. Balanced design		
A. $Y_{ij} = \beta_0 + U_{0j} + E_{ij}$	8.694	2.271
B. $Y_{ij} = \beta_0 + \beta_1 \bar{X}_j + U_{0j} + E_{ij}$	8.694	0.819
C. $Y_{ij} = \beta_0 + \beta_2(X_{ij} - \bar{X}_j) + U_{0j} + E_{ij}$	6.973	2.443
II. Unbalanced design		
A. $Y_{ij} = \beta_0 + U_{0j} + E_{ij}$	7.653	2.798
B. $Y_{ij} = \beta_0 + \beta_1 \bar{X}_j + U_{0j} + E_{ij}$	7.685	2.038
C. $Y_{ij} = \beta_0 + \beta_2(X_{ij} - \bar{X}_j) + U_{0j} + E_{ij}$	6.668	2.891

From Table 7.1 we see that in the balanced as well as in the unbalanced case, $\hat{\tau}_0^2$ increases as a within-group deviation variable is added as an explanatory variable to the model. Furthermore, for the balanced case, $\hat{\sigma}^2$ is not affected by adding a group-level variable to the model. In the unbalanced case, $\hat{\sigma}^2$ increases slightly when adding the group variable. When R^2 is defined as the proportional reduction in the residual variance parameters, as discussed above, then R^2 on the group level is negative for Model C, while for the entire data set R^2 on the pupil level is negative for Model B. Estimating σ^2 and τ_0^2 using the REML method results in slightly different parameter estimates. The pattern, however,

remains the same. It is argued below that defining R^2 as the proportional reduction in residual variance parameters $\hat{\sigma}^2$ and $\hat{\tau}_0^2$, respectively, is not the best way to define a measure analogous to R^2 in the linear regression model; and that the problems mentioned can be solved by using other definitions, leading to the measure denoted below by R_1^2 .

7.1.2 Definitions of the proportion of explained variance in two-level models

In multiple linear regression, the customary R^2 parameter can be introduced in several ways: for example, as the maximal squared correlation coefficient between the dependent variable and some linear combination of the predictor variables, or as the proportional reduction in the residual variance parameter due to the joint predictor variables. A very appealing principle to define measures of modeled (or explained) variation is the principle of *proportional reduction of prediction error*. This is one of the definitions of R^2 in multiple linear regression, and can be described as follows. A population of values is given for the explanatory and the dependent variables $(X_{1i}, \dots, X_{qi}, Y_i)$, with a known joint probability distribution; β is the value for which the expected squared error

$$\mathcal{E}\left(Y_i - \sum_{h=0}^q \beta_h X_{hi}\right)^2 \quad (7.1)$$

is minimal. (This is the definition of the ordinary least squares (OLS) estimation criterion. In this equation, β_0 is defined as the intercept and $X_{0i} = 1$ for all i .) If, for a certain case i , the values of X_{1i}, \dots, X_{qi} are unknown, then the best predictor for Y_i is its expectation $\mathcal{E}(Y_i)$, with mean squared prediction error $\text{var}(Y_i)$; if the values X_{1i}, \dots, X_{qi} are given, then the linear predictor of Y_i with minimum squared error is the regression value $\sum_h \beta_h X_{hi}$. The difference between the observed value Y_i and the predicted value $\sum_h \beta_h X_{hi}$ is the prediction error. Accordingly, the mean squared prediction error is defined as the value of (7.1) for the optimal (i.e., estimated) value of β .

The proportional reduction of the mean squared error of prediction is the same as the proportional reduction in the unexplained variance, due to the use of the variables X_1, \dots, X_q . Mathematically, it can be expressed as

$$R^2 = \frac{\text{var}(Y_i) - \text{var}(Y_i - \sum_h \beta_h X_{hi})}{\text{var}(Y_i)} = 1 - \frac{\text{var}(Y_i - \sum_h \beta_h X_{hi})}{\text{var}(Y_i)}.$$

This formula expresses one of the equivalent ways to define R^2 .

The same principle can be used to define ‘explained proportion of variance’ in the hierarchical linear model. For this model, however, there are several options with respect to what one wishes to predict. Let us consider a two-level model with dependent variable Y . In such a model, one can choose between predicting an individual value Y_{ij} at the lowest level, or a group mean \bar{Y}_j . On the basis of this distinction, two concepts of explained proportion of variance in a two-level model can be defined. The first, and most important, is the *proportional reduction of error for predicting an individual outcome*. The second is the *proportional reduction of error for predicting a group mean*. We treat the first concept here; the second is of less practical importance and is discussed in Snijders and Bosker (1994).

First consider a two-level random effects model with a random intercept and some predictor variables with fixed effects but no other random effects:

$$Y_{ij} = \gamma_0 + \sum_{h=1}^q \gamma_h X_{hij} + U_{0j} + R_{ij}. \quad (7.2)$$

Since we wish to discuss the definition of ‘explained proportion of variance’ as a population parameter, we assume temporarily that the vector γ of regression coefficients is known.

Explained variance at level one

We define the explained proportion of variance at level one. This is based on how well we can predict the outcome of Y_{ij} for a randomly drawn level-one unit i within a randomly drawn level-two unit j . If the values of the predictors X_{ij} are unknown, then the best predictor for Y_{ij} is its expectation; the associated mean squared prediction error is $\text{var}(Y_{ij})$. If the value of the predictor vector X_{ij} for the given unit is known, then the best linear predictor for Y_{ij} is the regression value $\sum_{h=0}^q \gamma_h X_{hij}$ (where X_{h0j} is defined as 1 for all h, j .) The associated mean squared prediction error is

$$\text{var}\left(Y_{ij} - \sum_h \gamma_h X_{hij}\right) = \sigma^2 + \tau_0^2.$$

The level-one explained proportion of variance is defined as the proportional reduction in mean squared prediction error:

$$R_1^2 = 1 - \frac{\text{var}(Y_{ij} - \sum_h \gamma_h X_{hij})}{\text{var}(Y_{ij})}. \quad (7.3)$$

Now let us proceed from the population to the data. The most straightforward way to estimate R_1^2 is to consider $\hat{\sigma}^2 + \hat{\tau}_0^2$ for the empty model,

$$Y_{ij} = \gamma_0 + U_{0j} + E_{ij}, \quad (7.4)$$

as well as for the fitted model (7.2), and compute 1 minus the ratio of these values. In other words, R_1^2 is just the proportional reduction in the value of $\hat{\sigma}^2 + \hat{\tau}_0^2$ due to including the X -variables in the model. For a sequence of nested models, the contributions to the estimated value of (7.3) due to adding new predictors can be considered to be the contribution of these predictors to the explained variance at level one.

To illustrate this, we once again use the data from the first (balanced) example, and estimate the proportional reduction of prediction error for a model where within-group and between-groups regression coefficients may be different.

From Table 7.2 we see that $\hat{\sigma}^2 + \hat{\tau}_0^2$ for model A amounts to 10.965, and for model D to 7.964. R_1^2 is thus estimated to be $1 - (7.964/10.965) = 0.274$.

Population values of R_1^2 are nonnegative

What happens to R_1^2 when predictor variables are added to the multilevel model? Is it possible that adding predictor variables leads to smaller values of R_1^2 ? Can we even be sure at all that these quantities are positive?

Table 7.2: Estimating the level-one explained variance (balanced data).

	$\hat{\sigma}^2$	$\hat{\tau}_0^2$
A. $Y_{ij} = \beta_0 + U_{0j} + E_{ij}$	8.694	2.271
D. $Y_{ij} = \beta_0 + \beta_1(X_{ij} - \bar{X}_j) + \beta_2\bar{X}_j + U_{0j} + E_{ij}$	6.973	0.991

It turns out that a distinction must be made between the population parameter R_1^2 and its estimates from data. *Population* values of R_1^2 in correctly specified models, with a constant group size n , become smaller when predictor variables are deleted, provided that the variables U_{0j} and E_{ij} on the one hand are uncorrelated with all the X_{ij} variables on the other hand (the usual model assumption).

For *estimates* of R_1^2 , however, the situation is different: these estimates sometimes do increase when predictor variables are deleted. When it is observed that an estimated value for R_1^2 becomes smaller by the addition of a predictor variable, or larger by the deletion of a predictor variable, there are two possibilities: either this is a chance fluctuation, or the larger model is misspecified. Whether the first or the second possibility is more likely will depend on the size of the change in R_1^2 , and on the subject-matter insight of the researcher. In this sense changes in R_1^2 in the ‘wrong’ direction serve as a *diagnostic* for possible misspecification. This possibility of misspecification refers to the fixed part of the model, that is, the specification of the explanatory variables having fixed regression coefficients, and not to the random part of the model. We return to this in Section 10.2.3.

7.1.3 Explained variance in three-level models

In three-level random intercept models (Section 4.9), the residual variance, or mean squared prediction variance, is the sum of the variance components at the three levels, $\sigma^2 + \tau_0^2 + \varphi_0^2$. Accordingly, the level-one explained proportion of variance can be defined here as the proportional reduction in the sum of these three variance parameters.

Example 7.1 Variance in maths performance explained by IQ.

In Example 4.8, Table 4.5 exhibits the results of the empty model (Model 1) and a model in which IQ has a fixed effect (Model 2). The total variance in the empty model is $7.816 + 1.746 + 2.124 = 11.686$ while the total unexplained variance in Model 2 is $6.910 + 0.701 + 1.109 = 8.720$. Hence the level-one explained proportion of variance is $1 - (8.720/11.686) = 0.25$.

7.1.4 Explained variance in models with random slopes

The idea of using the proportional reduction in the prediction error for Y_{ij} and \bar{Y}_j , respectively, as the definitions of explained variance at either level, can be extended to two-level models with one or more random regression coefficients. The formulas to calculate R_1^2 can be found in Snijders and Bosker (1994). However, the estimated values for R_1^2 usually change only very little when random regression coefficients are included in the model. The reason is that this definition of explained variance is based on prediction of the dependent variable from *observed* variables, and the knowledge of the precise specification of the random part only gives minor changes to the quality of this prediction.

The formulas for estimating R_1^2 in models with random intercepts only are very easy. Estimating R_1^2 in models with random slopes is more tedious. The simplest possibility for estimating R_1^2 in random slope models is to re-estimate the models as random intercept models with the same fixed parts (omitting the random slopes), and use the resulting parameter estimates to calculate R_1^2 in the usual (simple) way for random intercept models. This will normally yield values that are very close to the values for the random slopes model.

Example 7.2 *Explained variance for language scores.*

In Table 5.4, a model was presented for the data set on language scores in elementary schools used throughout Chapters 4 and 5. When a random intercept model is fitted with the same fixed part, the estimated variance parameters are $\hat{\tau}_0^2 = 8.10$ for level two and $\hat{\sigma}^2 = 38.01$ for level one. For the empty model, Table 4.1 shows that the estimates are $\hat{\tau}_0^2 = 18.12$ and $\hat{\sigma}^2 = 62.85$. This implies that explained variance at level one is $1 - (38.01 + 8.10) / (62.85 + 18.12) = 0.43$. This is a quite high explained variance. It can be deduced from the variances in Table 4.4.2 that the main part of this is due to IQ.

7.2 Components of variance¹

The preceding section focused on the total amount of variance that can be explained by the explanatory variables. In these measures of explained variance, only the fixed effects contribute. It can also be theoretically illuminating to decompose the observed variance of Y into parts that correspond to the various constituents of the model. This is discussed in this section for a two-level model.

For the dependent variable Y , the level-one and level-two variances in the empty model (4.6) are denoted by σ_E^2 and τ_E^2 , respectively. The total variance of Y is therefore $\sigma_E^2 + \tau_E^2$, and the components of variance are the parts into which this quantity is split. The first split, obviously, is the split of $\sigma_E^2 + \tau_E^2$ into σ_E^2 and τ_E^2 , and was extensively discussed in our treatment of the intraclass correlation coefficient.

To obtain formulas for a further decomposition, it is necessary to be more specific about the distribution of the explanatory variables. It is usual in single-level as well as in multilevel regression analysis to condition on the values of the explanatory variables, that is, to consider those as given values. In this section, however, all explanatory variables are regarded as random variables with a given distribution.

7.2.1 Random intercept models

For the random intercept model, we divide the explanatory variables into level-one variables X and level-two variables Z . Deviating from the notation in other parts of this book, matrix notation is used, and X and Z denote vectors.

The explanatory variables X_1, \dots, X_p at level one are collected in the vector X with value X_{ij} for unit i in group j . It is assumed more specifically that X_{ij} can be decomposed into independent level-one and level-two parts,

$$X_{ij} = X_{ij}^W + X_j^B \quad (7.5)$$

¹This is a more advanced section which may be skipped by the reader.

(i.e., a kind of multivariate hierarchical linear model without a fixed part). The expectation is denoted by $\mathbb{E}X_{ij} = \mu_X$, the level-one covariance matrix is

$$\text{cov}(X_{ij}^W) = \Sigma_X^W,$$

and the level-two covariance matrix is

$$\text{cov}(X_j^B) = \Sigma_X^B.$$

This implies that the overall covariance matrix of X is the sum of these,

$$\text{cov}(X_{ij}) = \Sigma_X^W + \Sigma_X^B = \Sigma_X.$$

Further, the covariance matrix of the group average for a group of size n is

$$\text{cov}(\bar{X}_j) = \frac{1}{n} \Sigma_X^W + \Sigma_X^B.$$

It may be noted that this notation deviates slightly from the common split of X_{ij} into

$$X_{ij} = (X_{ij} - \bar{X}_j) + \bar{X}_j. \quad (7.6)$$

The split (7.5) is a population-based split, whereas the more usual split (7.6) is sample-based. In the notation used here, the covariance matrix of the within-group deviation variable is

$$\text{cov}(X_{ij} - \bar{X}_j) = \frac{n-1}{n} \Sigma_X^W,$$

while the covariance matrix of the group means is

$$\text{cov}(\bar{X}_j) = \frac{1}{n} \Sigma_X^W + \Sigma_X^B.$$

For the discussion in this section, the present notation is more convenient.

The split (7.5) is not a completely innocuous assumption. The independence between X_{ij}^W and X_j^B implies that the covariance matrix of the group means is at least as large² as $1/(n-1)$ times the within-group covariance matrix of X .

The vector of explanatory variables $Z = (Z_1, \dots, Z_q)$ at level two has value Z_j for group j . The vector of expectations of Z is denoted

$$\mathbb{E}Z_j = \mu_Z,$$

and the covariance matrix is

$$\text{cov}(Z_j) = \Sigma_Z.$$

In the random intercept model (4.8), denote the vector of regression coefficients of the X s by

$$\gamma_X = (\gamma_{10}, \dots, \gamma_{p0})',$$

²The word ‘large’ is meant here in the sense of the ordering of positive definite symmetric matrices.

and the vector of regression coefficients of the Zs by

$$\gamma_Z = (\gamma_{01}, \dots, \gamma_{0q})'.$$

Taking into account the stochastic nature of the explanatory variables then leads to the following expression for the variance of Y :

$$\begin{aligned}\text{var}(Y_{ij}) &= \gamma_X' \Sigma_X \gamma_X + \gamma_Z' \Sigma_Z \gamma_Z + \tau_0^2 + \sigma^2 \\ &= \gamma_X' \Sigma_X^W \gamma_X + \gamma_X' \Sigma_X^B \gamma_X + \gamma_Z' \Sigma_Z \gamma_Z + \tau_0^2 + \sigma^2.\end{aligned}\quad (7.7)$$

This equation only holds if Z and \bar{X} are uncorrelated. For the special case where all explanatory variables are uncorrelated, this expression is equal to

$$\text{var}(Y_{ij}) = \sum_{h=1}^p \gamma_{h0}^2 \text{var}(X_h) + \sum_{h=1}^q \gamma_{0h}^2 \text{var}(Z_h) + \tau_0^2 + \sigma^2.$$

(This holds, for example, if there is only one level-one and only one level-two explanatory variable.) This formula shows that, in this special case, the contribution of each explanatory variable to the variance of the dependent variable is given by the product of the regression coefficient and the variance of the explanatory variable.

The decomposition of X into independent level-one and level-two parts allows us to indicate precisely which parts of (7.7) correspond to the unconditional level-one variance σ_E^2 of Y , and which parts to the unconditional level-two variance τ_E^2 :

$$\begin{aligned}\sigma_E^2 &= \gamma_X' \Sigma_X^W \gamma_X + \sigma^2, \\ \tau_E^2 &= \gamma_X' \Sigma_X^B \gamma_X + \gamma_Z' \Sigma_Z \gamma_Z + \tau_0^2.\end{aligned}$$

This shows how the within-group variation of the level-one variables eats up some part of the unconditional level-one variance; parts of the level-two variance are eaten up by the variation of the level-two variables, and also by the between-group (composition) variation of the level-one variables. Recall, however, the definition of Σ_X^B , which implies that the between-group variation of X is taken net of the ‘random’ variation of the group mean, which may be expected given the within-group variation of X_{ij} .

7.2.2 Random slope models

For the hierarchical linear model in its general specification given by (5.12), a decomposition of the variance is very complicated because of the presence of the cross-level interactions. Therefore the decomposition of the variance is discussed for random slope models in the formulation (5.15), repeated here as

$$Y_{ij} = \gamma_0 + \sum_{h=1}^q \gamma_h x_{hij} + U_{0j} + \sum_{h=1}^p U_{hj} x_{hij} + R_{ij}, \quad (7.8)$$

without bothering about whether some of the x_{hij} are level-one or level-two variables, or products of a level-one and a level-two variable.

Recall that in this section the explanatory variables X are stochastic. The vector $X = (X_1, \dots, X_q)$ of all explanatory variables has mean $\mu_{X(q)}$ and covariance matrix $\Sigma_{X(q)}$. The

subvector (X_1, \dots, X_p) of variables that have random slopes has mean $\mu_{X(p)}$ and covariance matrix $\Sigma_{X(p)}$. These covariance matrices could be split into within-group and between-group parts, but that is left to the reader.

The covariance matrix of the random slopes (U_{1j}, \dots, U_{pj}) is denoted by T_{11} and the $p \times 1$ vector of the intercept–slope covariances is denoted by T_{10} .

With these specifications, the variance of the dependent variable can be shown to be given by

$$\begin{aligned} \text{var}(Y_{ij}) &= \gamma' \Sigma_{X(q)} \gamma + \tau_0^2 + 2\mu'_{X(q)} T_{10} + \mu'_{X(q)} T_{11} \mu_{X(q)} \\ &\quad + \text{trace}(T_{11} \Sigma_{X(p)}) + \sigma^2. \end{aligned} \quad (7.9)$$

(A similar expression, but without taking the fixed effects into account, is given by Snijders and Bosker (1993) as formula (21).) A brief discussion of all terms in this expression is as follows.

1. The first term, $\gamma' \Sigma_{X(q)} \gamma$, gives the contribution of the fixed effects and may be regarded as the ‘explained part’ of the variance. This term could be split into a level-one and a level-two part as in the preceding subsection.

2. The next few terms,

$$\tau_0^2 + \mu'_{X(q)} T_{10} + \mu'_{X(q)} T_{11} \mu_{X(q)}, \quad (7.10)$$

should be seen as one piece. One could rescale all variables with random slopes to have a zero mean (cf. the discussion in Section 5.1.2); this would lead to $\mu_{X(q)} = 0$ and leave of this piece only the intercept variance τ_0^2 . In other words, (7.10) is just the intercept variance after subtracting the mean from all variables with random slopes.

3. The penultimate term, $\text{trace}(T_{11} \Sigma_{X(p)})$, is the contribution of the random slopes to the variance of Y . In the extreme case where all variables X_1, \dots, X_q would be uncorrelated and have unit variances, this expression reduces to the sum of squared random slope variances. This term also could be split into a level-one and a level-two part.

4. Finally, σ^2 is the residual level-one variability that can neither be explained on the basis of the fixed effects, nor on the basis of the latent group characteristics that are represented by the random intercept and slopes.

7.3 Glommary

Estimated variance parameters. These may go up when variables are added to the hierarchical linear model. This seems strange but it is a known property of the hierarchical linear model, indicating that one should be careful with the interpretation of the fine details of how the variance in the dependent variable is partitioned across the levels of the nesting structure.

Explained variance. Denoted by R^2_1 , this was defined as the proportional reduction in mean squared error for predicting the dependent variable, due to the knowledge of the

values of the explanatory variables. This was elaborated for two-level and three-level random intercept models.

Explained variance in random slope models. In this approach, this is so little different from R^2_1 in random intercept models that it is better for this definition of explained variance to pay no attention to the distinction between random slope and random intercept models.

Components of the variance of the dependent variable. These show how not only the fixed effects of explanatory variables, reflected in R^2_1 , but also the random effects contribute to the variance of the dependent variable.

8

Heteroscedasticity

The hierarchical linear model is quite a flexible model, and it has some other features in addition to the possibility of representing a nested data structure. One of these features is the possibility of representing multilevel as well as single-level regression models where the residual variance is not constant.

In ordinary least squares regression analysis, a standard assumption is *homoscedasticity*: residual variance is constant, that is, it does not depend on the explanatory variables. This assumption was made in the preceding chapters, for example, for the residual variance at level one and for the intercept variance at level two. The techniques used in the hierarchical linear model allow us to relax this assumption and replace it with the weaker assumption that variances depend linearly or quadratically on explanatory variables. This opens up an important special case of heteroscedastic models, that is, models with heterogeneous variances: heteroscedasticity where the variance depends on given explanatory variables. More and more programs implementing the hierarchical linear model also allow this feature. This chapter treats a two-level model, but the techniques treated (and the software mentioned) can be used also for heteroscedastic single-level regression models.

OVERVIEW OF THE CHAPTER

The chapter mainly treats heteroscedasticity at level one. First, the case is treated of the variance depending linearly on an explanatory variable. Next, the case of a quadratic dependence is treated, which also explains the representation of heteroscedasticity (linear as well as quadratic) as a direct extension of the hierarchical linear model of Chapter 5. Finally, a brief treatment is given of heteroscedasticity at level two; this follows the same principles as level-one heteroscedasticity.

8.1 Heteroscedasticity at level one

8.1.1 Linear variance functions

In a hierarchical linear model it sometimes makes sense to consider the possibility that the residual variance at level one depends on one of the predictor variables. An example

is a situation where two measurement instruments have been used, each with a different precision, resulting in two different values for the measurement error variance which is a component of the level-one variance. A linear dependence of the level-one residual variance on some variable X_1 can be expressed by

$$\text{level-one variance} = \sigma_0^2 + 2\sigma_{01}x_{1ij}, \quad (8.1)$$

where the value of X_1 for a given unit is denoted by x_{1ij} while the random part at level one now has two parameters, σ_0^2 and σ_{01} . The reason for incorporating the factor 2 and calling the parameter σ_{01} will become clear later, when also quadratic variance functions are considered. For example, when X_1 is a dummy variable with values 0 and 1, the residual variance is σ_0^2 for the units with $X_1 = 0$ and $\sigma_0^2 + 2\sigma_{01}$ for the units with $X_1 = 1$. When the level-one variance depends on more than variable, their effects can be added to the variance function (8.1) by adding terms $2\sigma_{02}x_{2ij}$, etc.

Example 8.1 Residual variance depending on gender.

In the example used in Chapters 4 and 5, the residual variance might depend on the pupil's gender. To investigate this in a model that is not overly complicated, we take the model of Table 5.4 and add the effect of gender (a dummy variable which is 0 for boys and 1 for girls). Table 8.1 presents estimates for two models: one with constant residual variances, and one with residual variances depending on gender.

Thus, Model 1 is a homoscedastic model and Model 2 a gender-dependent heteroscedastic model. The girls do much better on the language test: for the fixed effect of gender, Model 1 has $t = 2.407/0.201 = 12.0$, $p < 0.0001$. According to formula (8.1), the residual variance in Model 2 is 37.85 for boys and $37.85 - 2 \times 1.89 = 34.07$ for girls. The residual variance estimated in the homoscedastic Model 1 is very close to the average of these two figures. This is natural, since about half of the pupils are girls and half are boys. The difference between the two variances is significant: the deviance test yields $\chi^2 = 24,486.8 - 24,482.2 = 4.6$, $df = 1$, $p < 0.05$. However, the difference is not large. The conclusion is that, controlling for IQ, SES, the school means of these variables, and their interactions, girls score on average 2.4 higher than boys, and the results for girls are slightly less variable than for boys.

Analogous to the dependence due to the multilevel nesting structure as discussed in Chapter 2, heteroscedasticity has two faces: it can be a nuisance and it can be interesting. It can be a nuisance because the failure to take it into account may lead to a misspecified model and, hence, incorrect parameter estimates and standard errors. On the other hand, it can also be an interesting phenomenon in itself. When high values on some variable X_1 are associated with a higher residual variance, this means that for the units who score high on X_1 there is, within the context of the model being considered, more uncertainty about their value on the dependent variable Y . Thus, it may be interesting to look for explanatory variables that differentiate especially between units who score high on X_1 . Sometimes a nonlinear function of X_1 , or an interaction involving X_1 , could play such a role.

Example 8.2 Heteroscedasticity related to IQ.

Continuing the previous example, it is now investigated whether residual variance depends on IQ. The corresponding parameter estimates are presented as Model 3 in Table 8.2.

Comparing the deviance to Model 1 shows that there is a quite significant heteroscedasticity associated with IQ: $\chi^2 = 24,486.8 - 24,430.2 = 56.6$, $df = 1$, $p < 0.0001$. The level-one variance function is (cf. (8.1))

$$36.38 - 3.38 \text{ IQ}.$$

Table 8.1: Estimates for homoscedastic and heteroscedastic models.

	Model 1		Model 2	
Fixed effect	Coefficient	S.E.	Coefficient	S.E.
Intercept	40.426	0.265	40.435	0.266
IQ	2.249	0.062	2.245	0.062
SES	0.171	0.011	0.171	0.011
$\text{IQ} \times \text{SES}$	-0.020	0.005	-0.019	0.005
Gender	2.407	0.201	2.404	0.201
$\overline{\text{IQ}}$	0.769	0.293	0.749	0.292
$\overline{\text{SES}}$	-0.093	0.042	-0.091	0.042
$\overline{\text{IQ}} \times \overline{\text{SES}}$	-0.105	0.033	-0.107	0.033
Random part	Parameters	S.E.	Parameters	S.E.
<i>Level-two random part:</i>				
Intercept variance	8.321	1.036	8.264	1.030
IQ slope variance	0.146	0.065	0.146	0.065
Intercept–IQ slope covariance	-0.898	0.197	-0.906	0.197
<i>Level-one variance:</i>				
σ_0^2 constant term	35.995	0.874	37.851	1.280
σ_{01} gender effect			-1.887	0.871
Deviance	24,486.8		24,482.2	

This shows that language scores of the less intelligent pupils are more variable than language scores of the more intelligent. The standard deviation of IQ is 2.04 and the mean in this data set is 0.04. Thus, the range of the level-one variance, when IQ is in the range of the mean \pm twice the standard deviation, is between $36.38 - 3.38 \times (0.04 + 2 \times 2.04) = 22.45$ and $36.38 - 3.38 \times (0.04 - 2 \times 2.04) = 50.04$. This is an appreciable variation around the average value of 36.00 estimated in the homoscedastic Model 1.

Prompted by the IQ-dependent heteroscedasticity, the data were explored for effects that might differentiate between the pupils with lower IQ scores. Nonlinear effects of IQ and some interactions involving IQ were tried. It appeared that a nonlinear effect of IQ is discernible, represented better by a so-called spline function¹ than by a polynomial function of IQ. Specifically, the coefficient of the square of IQ turned out to be different for negative than for positive IQ values (recall that IQ was standardized to have an average of 0, which turned out to change to 0.04 after dropping a few cases with missing values). This is represented in Model 4 of Table 8.2 by introducing two nonlinear transformations of IQ: IQ_-^2 , which is the square but only for negative values and 0 elsewhere; and IQ_+^2 , being the square only for positive values and 0 elsewhere. The cutoff point of 0 was tried because it is a natural value, corresponding closely to the average; some other values also were tried

¹Spline functions (introduced more extensively in Section 15.2.2 and treated more fully, for example, in Fox, 2008, Chapter 17) are a more flexible class of functions than polynomials. They are polynomials of which the coefficients may be different on several intervals.

Table 8.2: Heteroscedastic models depending on IQ.

	Model 3		Model 4	
Fixed effect	Coefficient	S.E.	Coefficient	S.E.
Intercept	40.51	0.26	40.51	0.27
IQ	2.200	0.058	3.046	0.125
SES	0.175	0.011	0.168	0.011
$\text{IQ} \times \text{SES}$	-0.022	0.005	-0.016	0.005
Gender	2.311	0.198	2.252	0.196
$\overline{\text{IQ}}$	0.685	0.289	0.800	0.284
$\overline{\text{SES}}$	-0.087	0.041	-0.083	0.041
$\overline{\text{IQ}} \times \overline{\text{SES}}$	-0.107	0.033	-0.089	0.032
IQ_-^2			0.193	0.038
IQ_+^2			-0.260	0.033
Random part	Parameter	S.E.	Parameter	S.E.
<i>Level-two random effects:</i>				
Intercept variance	8.208	1.029	7.989	1.002
IQ slope variance	0.108	0.057	0.044	0.048
Intercept-IQ slope covariance	-0.733	0.187	-0.678	0.171
<i>Level-one variance parameters:</i>				
σ_0^2 constant term	36.382	0.894	36.139	0.887
σ_{01} IQ effect	-1.689	0.200	-1.769	0.191
Deviance	24,430.2		24,369.0	

and did not yield clearly better results. Thus, we use the definitions

$$\begin{aligned} \text{IQ}_-^2 &= \begin{cases} \text{IQ}^2 & \text{if } \text{IQ} < 0 \\ 0 & \text{if } \text{IQ} \geq 0, \end{cases} \\ \text{IQ}_+^2 &= \begin{cases} 0 & \text{if } \text{IQ} < 0 \\ \text{IQ}^2 & \text{if } \text{IQ} \geq 0. \end{cases} \end{aligned} \quad (8.2)$$

Adding these two variables to the fixed part gives a quite significant decrease in the deviance ($24,430.2 - 24,369.0 = 61.2$ with two degrees of freedom) and strongly reduces the random slope of IQ. Comparing the deviance with the model in which the random slope at level two of IQ is left out shows, however, that this random slope still is significant. The total fixed effect of IQ is now given by

$$\text{effect of IQ} = \begin{cases} 3.046 \text{IQ} + 0.193 \text{IQ}^2 & \text{if } \text{IQ} < 0 \\ 3.046 \text{IQ} - 0.260 \text{IQ}^2 & \text{if } \text{IQ} \geq 0. \end{cases} \quad (8.3)$$

The graph of this effect is shown in Figure 8.1. It is an increasing function which flattens out for low and especially for high values of IQ in a way that cannot be well represented by a quadratic or cubic

function, but for which a quadratic spline is a good representation. Perhaps this can be interpreted as a bottom and ceiling effect of the test used.

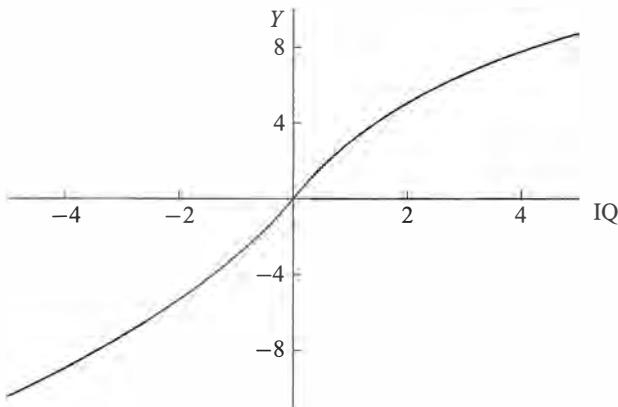


Figure 8.1: Effect of IQ on language test.

This is an interesting turn of this modeling exercise, and a nuisance only because it indicates once again that explanation of school performance is a quite complicated subject. Our interpretation of this data set toward the end of Chapter 5 was that there is a random slope of IQ (i.e. schools differ in the effect of IQ), with a variance estimated at 0.164 (a standard deviation of 0.40). Now it turns out that the data are clearly better described by a model in which IQ has a nonlinear effect, stronger in the middle range than toward its extreme values; in which the effect of IQ varies across schools, but with a variance of 0.044, corresponding to a standard deviation of 0.21, much smaller than the earlier value; and in which the combined effects of IQ, SES, their school averages and interactions, and gender predict the language scores for high-IQ pupils much better than for low-IQ pupils. Thus, a considerable portion of what seemed to be between-school differences in the effect of IQ on the language test now turns out to be partly the consequence of nonlinear effects of IQ and heteroscedasticity.

8.1.2 Quadratic variance functions

The formal representation of level-one heteroscedasticity is based on including random effects at level one, as spelled out in Goldstein (2011) – see the remarks in Section 3.1 of Goldstein’s book on a complex random part at level one. In Section 5.1.1 it was remarked that random slopes (i.e., random effects at level two) lead to heteroscedasticity. This also holds for random effects at level one. Consider a two-level model and suppose that the level-one random part is

$$\text{random part at level one} = R_{0ij} + R_{1ij}x_{1ij}. \quad (8.4)$$

Denote the variances of R_{0ij} and R_{1ij} by σ_0^2 and σ_1^2 , respectively, and their covariance by σ_{01} . The rules for calculating with variances and covariances imply that

$$\text{var}(R_{0ij} + R_{1ij}x_{1ij}) = \sigma_0^2 + 2\sigma_{01}x_{1ij} + \sigma_1^2x_{1ij}^2. \quad (8.5)$$

Formula (8.4) is just a formal representation, useful for specifying this heteroscedastic model. For the interpretation one should look not at (8.4) with its would-be random effect,

but rather at only the right-hand side of (8.5). This formula can be used without the interpretation that σ_0^2 and σ_1^2 are variances and σ_{01} a covariance; these parameters might be any numbers. The formula only implies that the residual variance is a quadratic function of x_{1ij} . In the previous section, the case was used where $\sigma_1^2 = 0$, producing the linear function (8.1). If a quadratic function is desired, all three parameters are estimated from the data.

Example 8.3 Educational level attained by pupils.

This example is about a cohort of pupils entering secondary school in 1989, studied by Dekkers et al. (2000). The question is how well the educational level attained in 1995 can be predicted from individual characteristics and school achievement at the end of primary school. The data is about 13,007 pupils in 369 secondary schools. The dependent variable is an educational attainment variable, defined as 12 minus the minimum number of additional years of schooling it would take theoretically for this pupil in 1995 to gain a certificate giving access to university. The range is from 4 to 13 (a value of 13 means that the pupil is already a first-year university student). Explanatory variables are teacher's rating at the end of primary school (indicating the most suitable type of secondary school, range from 1 to 4), achievement on three standardized tests (so-called CITO tests, on language, arithmetic, and information processing – all with a mean value of about 11 and a standard deviation between 4 and 5), socio-economic status (a discrete ordered scale with values from 1 to 6), gender (0 for boys, 1 for girls), and minority status (based on the parents' country of birth, 0 for the Netherlands and other industrialized countries, 1 for other countries).

Table 8.3 presents the results of a random intercept model (Model 1). Standard errors are quite small due to the large number of pupils in this data set. The explained proportion of variance at level one is $R_1^2 = 0.39$. Model 2 shows the results for a model where residual variance depends quadratically on SES. It follows from (8.5) that residual variance here is given by

$$\text{residual variance} = 3.475 - 0.632 \text{SES} + 0.056 \text{SES}^2.$$

The difference in deviance between Models 1 and 2 ($\chi^2 = 97.3$, $df = 2$, $p < 0.0001$) indicates that the dependence of residual variance on SES is quite significant. The variance function decreases curvilinearly from the value 2.91 for SES = 1 to a minimum value of 1.75 for SES = 6. This implies that when educational attainment is predicted by the variables in this model, the uncertainty in the prediction is highest for low-SES pupils. It is reassuring that the estimates and standard errors of other effects are not appreciably different between Models 1 and 2.

The specification of this random part was checked in the following way. First, the models with only a linear or only a quadratic variance term were estimated separately. This showed that both variance terms are significant. Further, it might be possible that the dependence of the residual variance on SES is a random slope in disguise. Therefore a model with a random slope (a random effect at level two) for SES also was fitted. This showed that the random slope was barely significant and not very large, and did not take away the heteroscedasticity effect.

The SES-dependent heteroscedasticity led to the consideration of nonlinear effects of SES and interaction effects involving SES. Since the SES variable assumes values 1 to 6, five dummy variables were used contrasting the respective SES values to the reference category SES = 3. In order to limit the number of variables, the interactions of SES were defined as interactions with the numerical SES variable rather than with the categorical variable represented by these dummies. For the same reason, for the interaction of SES with the CITO tests only the average of the three CITO tests (range from 1 to 20, mean 11.7) was considered. Product interactions of SES with gender, with the average CITO score, and with minority status were considered. As factors in the product, SES and CITO tests were centered approximately by using (SES – 3) and (CITO average – 12). Gender and minority status, being 0–1 variables, did not need to be centered.

This implies that although SES is represented by dummies (i.e., as a categorical variable) in the main effect, it is used as a numerical variable in the interaction effects. (The main effect of SES as

Table 8.3: Heteroscedastic model depending quadratically on SES.

	Model 1		Model 2	
Fixed effect	Coefficient	S.E.	Coefficient	S.E.
Intercept	6.021	0.066	6.024	0.067
Teacher's rating	0.490	0.023	0.487	0.023
CITO Arithmetic	0.0567	0.0042	0.0578	0.0042
CITO Information	0.0651	0.0049	0.0648	0.0049
CITO Language	0.0361	0.0049	0.0365	0.0048
SES	0.168	0.013	0.166	0.013
Gender	0.322	0.028	0.318	0.027
Random part	Parameter	S.E.	Parameter	S.E.
<i>Level-two random effect:</i>				
Intercept variance	0.128	0.014	0.124	0.014
<i>Level-one variance parameters:</i>				
σ_0^2 constant term	1.982	0.025	3.475	0.247
σ_{01} linear SES effect			-0.316	0.064
σ_1^2 quadratic SES effect			0.056	0.016
Deviance	46,205.1		46107.8	

a numerical variable is implicitly also included, because it can be represented also as an effect of the dummy variables. Therefore the numerical SES variable does not need to be added to the model.)

Minority status does not have a significant main effect when added to Model 2, but the main effect was added to facilitate interpretation of the interaction effect. The results are in Table 8.4.

Model 3 as a whole is a strong improvement over Model 2: the difference in deviance is $\chi^2 = 44.8$ for $df = 7$ ($p < 0.0001$). For the evaluation of the nonlinear effect of SES, note that since SES = 3 is the reference category, the parameter value for SES = 3 should be taken as 0.0. This demonstrates that the effect of SES is nonlinear but it is indeed an increasing function of SES. The differences between the SES values 1, 2, and 3 are larger than those between the values 3, 4, 5, and 6. The interaction effects of SES with gender and with minority status are significant. The main effect of minority status corresponds with its effect for SES = 3, since the product variable was defined using SES - 3, and is practically nil. Thus, it turns out that, when the other included variables (including the main effect of SES) are controlled for, pupils with parents from a nonindustrialized country attain a higher educational level than those with parents from the Netherlands or another industrialized country when they come from a low-SES family, but a lower level if they come from a high-SES family.

In Model 3, residual variance is

$$\text{residual variance} = 3.422 - 0.604 \text{ SES} + 0.053 \text{ SES}^2.$$

This decreases from 2.87 for SES = 1 to 1.71 for SES = 6. Thus, with the inclusion of interactions and a nonlinear SES effect, residual variance has hardly decreased.

Table 8.4: Heteroscedastic model with interaction effects.

Model 3		
Fixed effect	Coefficient	S.E.
Intercept	6.594	0.069
Teacher's rating	0.484	0.023
CITO Arithmetic	0.0564	0.0042
CITO Information	0.0632	0.0050
CITO Language	0.0355	0.0049
SES = 1	-0.827	0.155
SES = 2	-0.390	0.050
SES = 4	0.133	0.034
SES = 5	0.389	0.047
SES = 6	0.500	0.070
Gender	0.384	0.034
Minority	-0.005	0.064
SES × Gender	-0.074	0.023
SES × CITO	0.0056	0.0035
SES × Minority	-0.219	0.050
Random part	Parameter	S.E.
<i>Level-two random effect:</i>		
Intercept variance	0.119	0.014
<i>Level-one variance parameters:</i>		
σ_0^2 constant term	3.422	0.246
σ_{01} linear SES effect	-0.302	0.064
σ_1^2 quadratic SES effect	0.053	0.016
Deviance	46,063.0	

Model 3 was obtained, in the search for heteroscedasticity, in a data-driven rather than theory-driven way. Therefore one may question the validity of the tests for the newly included effects: are these not the result of capitalization on chance? The interaction effect of SES with minority status has such a high t -value, 4.4, that its significance is beyond doubt, even when the data-driven selection is taken into account. For the interaction of SES with gender this is less clear. For convincing hypothesis tests, it would have been preferable to use cross-validation: split the data into two subsets and use one subset for the model selection and the other for the test of the effects. Since the two subsets should be independent, it would be best to select half the schools at random and use all pupils in these schools for one subset, and the other schools and their pupils for the other.

More generally, the residual variance may depend on more than one variable; in terms of representation (8.4), several variables may have random effects at level one. These can be level-two as well as level-one variables. If the random part at level one is given by

$$\text{random part at level one} = R_{0ij} + R_{1ij}x_{1ij} + \dots + R_{pij}x_{pij},$$

while the variances and covariances of the R_{hij} are denoted by σ_h^2 and σ_{hk} , then the variance function is

$$\text{residual variance} = \sum_{h=0}^p \sigma_h^2 x_{hj}^2 + 2 \sum_{h=0}^{p-1} \sum_{k=h+1}^p \sigma_{hk} x_{hij} x_{kij}. \quad (8.6)$$

This complex level-one variance function can be used for any values for the parameters σ_h^2 and σ_{hk} , provided that the residual variance is positive. The simplest case is to include only σ_0^2 and the ‘covariance’ parameters σ_{0h} , leading to the linear variance function

$$\text{residual variance} = \sigma_0^2 + 2\sigma_{01}x_{1ij} + 2\sigma_{02}x_{2ij} + \dots + 2\sigma_{0p}x_{pij}.$$

Correlates of diversity

It may be important to investigate the factors that are associated with outcome variability. For example, Raudenbush and Bryk (1987) (see also Raudenbush and Bryk, 2002, pp. 131–134) investigated the effects of school policy and organization on mathematics achievement of pupils. They did this by considering within-school dispersion as a dependent variable. The preceding section offers an alternative approach which remains closer to the hierarchical linear model. In this approach, relevant level-two variables are considered as potentially being associated with level-one heteroscedasticity.

Example 8.4 School composition and outcome variability.

Continuing the preceding example on educational attainment predicted from data available at the end of primary school, it is now investigated whether composition of the school with respect to socio-economic status is associated with diversity in later educational attainment. It turns out that socio-economic status has an intraclass correlation of 0.25, which is quite high. Therefore the average socio-economic status of schools could be an important factor in the within-school processes associated with average outcomes but also with outcome diversity.

To investigate this, the school average of SES was added to Model 3 of Table 8.4 both as a fixed effect and as a linear effect on the level-one variance. The nonsignificant SES-by-CITO interaction was deleted from the model. The school average of SES ranges from 1.4 to 5.5, with mean 3.7 and standard deviation 0.59. This variable is denoted by SA-SES. Its fixed effect is 0.417 (S.E. 0.109, $t = 3.8$). We further present only the random part of the resulting model in Table 8.5.

To test the effect of SA-SES on the level-one variance, the model was estimated also without this effect. This yielded a deviance of 46,029.6, so the test statistic is $\chi^2 = 46,029.6 - 45,893.0 = 136.6$ with $df = 1$, which is very significant. The quadratic effect of SA-SES was estimated both as a main effect and for level-one heteroscedasticity, but neither was significant.

How important is the effect of SA-SES on the level-one variance? The standard deviation of SA-SES is 0.59, so four times the standard deviation (the difference between the few percent highest-SA-SES and the few percent lowest-SA-SES schools) leads to a difference in the residual variance of $4 \times 0.59 \times 0.265 = 0.63$. For an average residual level-one variance of 2.0 (see Model 1 in Table 8.3), this is an appreciable difference.

This ‘random effect at level one’ of SA-SES might be explained by interactions between SA-SES and pupil-level variables. The interactions of SA-SES with gender and with minority status were considered. Adding these to Model 4 yielded interaction effects of -0.219 (S.E. 0.096, $t = -2.28$) for SA-SES by minority status and -0.225 (S.E. 0.050, $t = 4.50$) for SA-SES by gender. This implies that, although a high school average for SES leads to higher educational attainment on average (the main effect of 0.417 reported above), this effect is weaker for minority pupils and for girls. These

Table 8.5: Heteroscedastic model depending on average SES.

Model 4		
Random effect	Parameter	S.E.
<i>Level-two random effect:</i>		
Intercept variance	0.101	0.013
<i>Level-one variance parameters:</i>		
σ_0^2 constant term	5.203	0.282
σ_{01} linear SES effect	-0.331	0.063
σ_1^2 quadratic SES effect	0.078	0.022
σ_{02} linear SA-SES effect	-0.265	0.022
Deviance	45,893.0	

interactions did, however, not lead to a noticeably lower effect of SA-SES on the residual level-one variability.

8.2 Heteroscedasticity at level two

For the intercept variance and the random slope variance in the hierarchical linear model it was assumed in preceding chapters that they are constant across groups. This is a homoscedasticity assumption at level two. If there are theoretical or empirical reasons to drop this assumption, it could be replaced by the weaker assumption that these variances depend on some level-two variable Z . For example, if Z is a dummy variable with values 0 and 1, distinguishing two types of groups, the assumption would be that the intercept and slope variances depend on the group. In this section we only discuss the case of level-two variances depending on a single variable Z ; this discussion can be extended to variances depending on more than one level-two variable.

Consider a random intercept model in which one assumes that the intercept variance depends linearly or quadratically on a variable Z . The intercept variance can then be expressed by

$$\text{intercept variance} = \tau_0^2 + 2\tau_{01}z_j + \tau_1^2 z_j^2, \quad (8.7)$$

for parameters τ_0^2 , τ_{01} , and τ_1^2 . For example, in organizational research, when the level-two units are organizations, it is possible that small-sized organizations are (because of greater specialization or other factors) more different from one another than large-sized organizations. Then Z could be some measure for the size of the organization, and (8.7) would indicate that it depends on Z whether differences between organizations tend to be small or large; where ‘differences’ refer to the intercepts in the multilevel model. Expression (8.7) is a quadratic function of z_j , so that it can represent a curvilinear dependence of the intercept

variance on Z . If a linear function is used (i.e., $\tau_1^2 = 0$), the intercept variance is either increasing or decreasing over the whole range of Z .

Analogous to (8.4), this variance function can be obtained by using the ‘random part’

$$\text{random part at level two} = U_{0j} + U_{1j}z_j. \quad (8.8)$$

Thus, strange as it may sound, the level-two variable Z formally gets a random slope at level two. (Recall that in Section 5.1.1 it was observed that random slopes for level-one variables also create some kind of heteroscedasticity, namely, heteroscedasticity of the observations Y_j .)

The parameters τ_0^2 , τ_1^2 , and τ_{01} are, as in the preceding section, not to be interpreted themselves as variances and a corresponding covariance. The interpretation is by means of the variance function (8.8). Therefore it is not required that $\tau_{01}^2 \leq \tau_0^2 \times \tau_1^2$. To put it another way, ‘correlations’ defined formally by $\tau_{01}/\sqrt{\tau_0\tau_1}$ may be larger than 1 or smaller than -1, even infinite, because the idea of a correlation does not make sense here. An example of this is provided by the linear variance function for which $\tau_1^2 = 0$ and only the parameters τ_0^2 and τ_{01} are used.

In a similar way, the random slope variance of a level-one variable X_1 can be made to depend on a level-two variable Z by giving both X_1 and the product $X_1 \times Z$ a random slope at level two. To avoid computational difficulties it is advisable to use a variable Z which has been centered, even if only very approximately.

8.3 Glommary

Heteroscedasticity. Residual variance depending on explanatory variables. The hierarchical linear model can be extended so that residual variances depend, linearly or quadratically, on explanatory variables. This can be a nuisance but also interesting.

Heteroscedasticity at level one. This is achieved by formally giving a variable a random effect at level one. This should not be regarded as a ‘real’ random effect but merely as a device to achieve a linear or quadratic variance function.

Correlates of diversity. Variables that are predictive of residual variances. This term is used in studies where the interest is in explaining variances rather than explaining means. This is a nice interpretation or theoretical framing of heteroscedastic models. When heteroscedasticity is found, this can lead to searching for variables that, when included as fixed effects in the model, might explain this heteroscedasticity – for example, nonlinear transformations of explanatory variables, or interactions.

Random slopes or heteroscedasticity? As a by-product of our example of heteroscedasticity at level one, we found that what in Chapter 5 appeared to be a random slope for IQ, could be better described in part as a consequence of a nonlinear effect of IQ in the fixed part (modeled by a spline function).

Heteroscedasticity at level two. This can be achieved by formally giving a variable a random effect at level two. Again, this is not a ‘real’ random effect but only a device to obtain the heteroscedasticity.

9

Missing Data

In scientific practice, available data often are incomplete – because research subjects did not participate fully, registers were not complete, an apparatus failed, or for whatever reason. For a long time, it was not uncommon for researchers to ignore the associated difficulties, and one usual approach was working with complete cases only and leaving out incomplete cases: a practice called *listwise deletion* or *complete case analysis*. Another usual approach was *available case analysis*, where each parameter is estimated using the data points directly relevant to this particular parameter; for example, the expected value of some variable is estimated by the mean of this variable over the cases where the variable was observed. It is now recognized that these approaches can lead to incorrect conclusions, and therefore it is no longer acceptable to use them unthinkingly as default methods. Appropriate ways have been developed to deal with missing data, and this chapter serves as an introduction to some of these methods and their application to multilevel analysis.

There are two basic reasons for attempting to deal with incomplete data in a suitable way. Most important, in many cases the absence of data is related to meaningful information, and ignoring this can bias the results obtained. Second, deleting incomplete cases is wasteful of information in the sense that standard errors of estimators are unnecessarily large and hypothesis tests have unnecessarily low power.

Classic texts on missing data analysis in general are Little and Rubin (2002) and Schafer (1997). Relatively brief and accessible overviews are Carpenter and Kenward (2008) and Graham (2009). A textbook on missing data in longitudinal data sets analyzed by hierarchical linear models is Molenberghs and Kenward (2007).

OVERVIEW OF THE CHAPTER

The chapter starts by outlining the basis of the modern theory of inference from incomplete data. This rests on the concept of ‘missingness at random’, which expresses that the fact itself that some data points are missing does not potentially reveal anything about the unobserved values that is not already contained in the observed data. Advice is given about potentially collecting auxiliary data that could make this assumption more reasonable. How to deal with missing data in situations where this assumption is not satisfied depends strongly on the particular case under study and the reasons why data points are missing; this is outside the scope of this book.

Of the various techniques for analysis of incomplete data, we focus on two that are suitable for multilevel analysis: maximum likelihood and multiple imputation. We give most attention to the approach of multiple imputation because it has great potential for general multilevel data structures, but requires some explanation if one is to understand it properly. Since this is a topic that for the special case of multilevel data is only beginning to be implemented in statistical software, we go somewhat deeper into the technical details of how to carry out this type of data analysis than we do in most other chapters. In particular, we treat in some detail the very flexible technique of multiple imputation by chained equations. We treat incompleteness for general multilevel data structures, without paying special attention to longitudinal data.

9.1 General issues for missing data

Suppose that we are working with an incomplete data set. To model such a data set, we need to think about a model for the hypothetical complete data set of which this is the currently observed part. Let us consider each explanatory and each dependent variable for each case in the data as a data point; in the terminology of Chapter 4, this is each variable y_{ij} , x_{pij} , and z_{qj} . Then any data point could be either observed or missing. (We suppose that we do have complete information about the nesting structure, that is, in a two-level design, for all level-one units we know which level-two unit they belong to. Cases where this kind of information is missing were discussed by Hill and Goldstein, 1998). The hypothetical complete data set has observed values for all data points, whereas the actual incomplete data set lacks some of these data points. In the model for the incomplete data, this can be indicated by *missingness indicators* defined for all data points, which convey the information as to whether a given data point is missing or observed, for example, by values 1 and 0, or true and false, reflecting missing versus observed.

Here we must note a distinction that is usually glossed over. In regression-type models, of which the hierarchical linear model is one example, where there is a distinction between dependent and explanatory (or independent) variables, the model is *conditional on* the explanatory variables. Mostly the latter variables are treated as fixed – they are whatever they are, and they might as well have been determined by the researcher, as they are in purely experimental designs. If, however, we wish to make sense of an incomplete data set in which some of the data points for explanatory variables x_{pij} and z_{qj} are missing, then we must entertain a model for the complete data set in which the explanatory variables are random, rather than fixed, variables, and in which they have some simultaneous probability distribution. This probability distribution, when it has been estimated, will allow us to say that some hypothetical values for the unobserved variables are more likely than others, and this kind of knowledge will be used in the analysis of the incomplete data set, for example, by attempting to fill in ('impute') the unobserved data points. To summarize, in the usual approach the explanatory variables are treated as if they have fixed, deterministic values; but in the analysis of an incomplete data set the variables that have some missing values are treated as random variables. If there are some variables that have no missing cases at all, these may be regarded as deterministic values.

The extra step taken in modern treatments of missing data is to regard the missingness indicators themselves also as random variables that are part of the data. This alerts us to the possibility that they may contain meaningful information; for example, if some pupils do

not show up for an exam because they expect to fail, the missingness is predictive of what would be a poor exam mark, if it had been observed. It is sometimes meaningful to regard the missingness variables as part of the substantively interesting social process that is being studied.

Rubin (1976) defined three types of missingness:

- *missingness completely at random (MCAR)*, where the missingness indicators are independent of the complete data. An example is randomly failing measurement apparatus.
- *missingness at random (MAR)*, where, conditionally given the observed data, the missingness indicators are independent of the unobserved data.¹ This means that the missingness itself contains no information about the unobserved values that is not also contained in the observed data.² An example is repeated measures of ill patients where initially all patients are being observed and observation stops at recovery, while recovery is recorded in the data.
- *missingness not at random (MNAR)*, where, conditionally given the observed data, the missingness indicators are dependent on the unobserved data. An example is repeated measures of ill patients where some patients do not show up for data collection if they feel very well or very ill, and the reason for absence is unknown.

The first situation, MCAR, is the pleasant situation in which we can almost forget about the incompleteness of the data, and throwing away the incomplete cases would entail only a loss of statistical efficiency (and therefore, loss of power) but no bias. The second, MAR, is the potentially difficult situation where naive approaches such as complete case analysis can be biased and hence misleading, but where it is possible to make judicious use of the observations and avoid bias. For the MCAR as well as the MAR case it may be necessary to use special statistical methods to avoid efficiency loss. By the way, the terminology is potentially misleading because many people might interpret the term ‘missingness at random’ as meaning what was defined above as ‘missingness completely at random’, but these are the generally employed terms so we have to use them.

The third situation, MNAR, leads to serious difficulties for the researcher because the missingness pattern contains information about the unobserved values, but in practice we do not know *which* information. We wish to say something about the world but we have observed too little of it. To draw conclusions from the data it is necessary in this case to make additional assumptions about how missingness is related to the values of the data points that would have been observed if they had not been missing. While the researcher must attempt to make plausible assumptions of this kind, they will not be completely testable. In other words, to analyze data in the MNAR case we need to make assumptions

¹The phrase ‘conditionally given the observed data’ refers to the assumption that the observed values are known, but without knowing that these variables are all that what is observed; thus, the conditioning is not on the missingness indicators.

²Formally, for some of what is asserted below about MAR, an extra condition is needed. This is that the parameters for the missingness process and those for the model explaining the primary dependent variable are distinct. This condition is called *separability*. It is necessary to ensure, together with MAR, that the missingness variables themselves are not informative for the primary research question. This will be tacitly assumed below when the MAR condition is considered. MAR and separability together define *ignorability* of the missingness mechanism.

going beyond the available data and beyond the model that we would use for analyzing complete data. To still say something in this case about the uncertainty of the conclusions drawn from the data, it will be helpful to carry out sensitivity analyses studying the sensitivity of the conclusions to the assumptions made. In the MNAR situation, any data analysis will leave bigger questions than in the MAR situation.

9.1.1 Implications for design

Even though this may seem like forcing an open door, we must emphasize that it is important in the phase of design and data collection to attempt to collect complete data, and if this is impossible, to minimize the amount of missing data. Second, if missing data are unavoidable, it will be helpful to record information about what led to the missingness (e.g., in the case of data about employees in an organization, if data for an employee are totally absent, whether it was because of illness, or refusal to cooperate, or negligence, or loss of the questionnaire, etc.) and about auxiliary variables that may be predictive of missingness and that can be also collected in case of missingness (continuing the preceding example, some of the reasons for nonresponse will be related with the general cooperative attitude of the employee, and the research design could include a question to supervisors to rate the cooperative attitude for each of their subordinates). It is also sometimes possible to observe an auxiliary variable that may serve as a rough proxy for the missing variable (which then should be regarded as a separate variable and collected, if possible, for all cases). In most social research, the process leading to incompleteness of collected data is itself a social process, and the research will gain by trying to understand it.

The analysis of an incomplete data set should start with a description of the number of missing data points, and how these are distributed across variables and across higher-level units. To understand the amount and patterns of missingness, it may be helpful to employ methods such as pilot studies, debriefing of respondents and those carrying out the data collection, and reflection on the data collection process. An explicit model of the missingness pattern of a given variable may be constructed by a multilevel logistic regression analysis of the binary missingness indicators as dependent variables (see Chapter 17). Thinking about such a model can elucidate the difference between MAR and MNAR: if MNAR holds then hypothetically knowing the unobserved data would help in predicting the missingness indicators, whereas it would not help under the MAR condition.

Most methods for handling incomplete data are based on the MAR assumption. This is because, on the one hand, MCAR is rather rare in the practice of social research. On the other hand, although MNAR may often be more realistic than MAR, it is much more difficult to handle than MAR; and, as suggested above, by collecting auxiliary information that is predictive of missingness the researcher may be able to push the research design from MNAR in the direction of MAR. Also, even in the MNAR case, a principled analysis employing the MAR assumption will mostly lead to more credible results than a naive way of handling the missing data.

9.2 Missing values of the dependent variable

For missingness of the dependent variable, MAR means in practice (under an additional assumption of mutual independence of missingness indicators) that the missingness of the

dependent variable may depend on observed explanatory variables for this particular level-one unit, but not on unobserved ones that are correlated with the dependent variable, nor on the unobserved value of the dependent variable itself. For example, if particularly high or low values of the dependent variable were more likely to be missing but this is not totally predictable based on the observed explanatory variables, then MNAR would hold. In this section it further is assumed that MAR holds.

Suppose that there are only missing data for the dependent variable, all other variables are fully observed, and the explanatory variables are considered as deterministic variables (so that we are not interested in estimating the distribution of the explanatory variables). If all observed variables are potentially included among the explanatory variables – this will depend on the research question – the situation is relatively uncomplicated. Since the hierarchical linear model, like other regression-type techniques, is conditional on the values for the explanatory variables, such units can safely be dropped from the data set for the basic analysis according to the hierarchical linear model (cf. Little and Rubin, 2002, p. 237).

However, the research question may be such that some observed variables X_{extra} are excluded from the model that is currently being estimated. This may be because of causality arguments, or because of the particular focus of the current research question, as will be discussed on p. 136. The variables X_{extra} might also include the auxiliary variables collected to explain the missingness as suggested in Section 9.1.1. Then it is possible that MAR holds when X_{extra} is included, but MNAR would hold if it were excluded from consideration. It also is possible that X_{extra} has a strong predictive value for the dependent variable. In such cases the variable X_{extra} should be included in the procedures for handling the missing data, even if it is not to be used further among the explanatory variables for the model that is currently being estimated.

For multivariate hierarchical linear models, that is, models with multiple dependent variables (Chapter 16), the situation is more complicated, because the patterns of missingness may well differ between the various dependent variables. The treatment of incomplete data for such models, with missing values on the dependent variables, was discussed by Schafer and Yucel (2002).

9.3 Full maximum likelihood

An ideal way to proceed under MAR is to formulate the likelihood of the observed data and find the corresponding maximum likelihood (ML) estimates. This is sometimes referred to as ‘full information maximum likelihood’ (FIML).

When there are missing values only on the dependent variable and missingness depends only on the explanatory variables, then the missingness mechanism is MAR. The preceding section argued that then one may simply compute the ML estimates of the regular hierarchical linear model in which the level-one units with a missing dependent variable are dropped. These are then the ML estimates for the observed data. This was also proposed for incomplete multivariate data, or incomplete longitudinal data with a fixed occasion design, on p. 257. One of the advantages of the algorithms available for the hierarchical linear model is that they easily allow FIML estimation for incomplete longitudinal or multivariate data.

If there are missing values also in explanatory variables and the missing data pattern is monotone, then the full maximum likelihood approach is feasible. ‘Monotone’ means that

the variables can be ordered such that the cases with missing values on a given variable are a subset of the cases with missing values on the next variable. For nonmonotone missingness patterns the full maximum likelihood approach is more cumbersome. This is also the case when the missingness in the dependent variable depends on auxiliary variables X_{extra} (as described above) that are observed but that the researcher does not wish to use among the explanatory variables. There are methods for parameter estimation in such situations by FIML (see Molenberghs and Kenward, 2008; Ibrahim and Molenberghs, 2009), but the multiple imputation method of Section 9.4 is more flexible and we restrict ourselves to presenting the latter method. Further discussion on the choice between FIML and multiple imputation is given by Graham (2009, p. 560).

9.4 Imputation

Imputation means filling in something (hopefully reasonable) for the missing data points, which then leads to a completed data set so that a regular complete-data analysis is possible. The general literature on missing data (e.g., Little and Rubin, 2002; Schafer and Graham, 2002) provides arguments and examples explaining why simple methods such as imputation by an overall mean, and also slightly more complicated methods such as imputation by a groupwise mean, will yield biased results and underestimation of standard errors. Rubin (1987) had the insight that *multiple stochastic imputation* is an approach which leads to good (approximately unbiased) results under MAR. This approach now is one of the major methods employed in the analysis of incomplete data, and an extensive discussion can be found, for example, in Graham (2009). Multiple stochastic imputation works in three steps:

1. Construct several completed data sets, in each of which the missing values are imputed by a random draw from the distribution of missing data, given the observed data. These random draws must be independent across the multiple data sets.
2. Use a regular complete-data analysis method for each of these multiple data sets, producing parameter estimates and standard errors for the hypothetical situation where these are real observed data sets.
3. Combine the parameter estimates and standard errors across these multiple analyses into a single set of results. The resulting overall standard errors will combine the within-data-set standard errors obtained in step 2 with the between-data-set variability of the estimates.

Note that this procedure does not require any type of knowledge of the model for the missingness variables beyond the validity of the MAR assumption. The virtue of this approach is that step 2 is normally straightforward, consisting, in our case, of the estimation of the hierarchical linear model or one of its variants. The combination formulas of step 3 are simple, as we shall see below. The catch is in step 1. This requires knowledge of the distribution of the missing data given the observed data, but this will depend on the parameters that we are trying to estimate in the whole process³ – the dog seems to bite its own tail. The

³At least it will depend on the parameters of the imputation model; not on parameters of the model of interest that are not included in the imputation model.

approach is still feasible because it is possible to let the dog bite its own tail *ad infinitum*, and cycle through the steps repeatedly until stability of the results is achieved. Especially in the Bayesian paradigm this can be formulated in a very elegant way as a kind of ‘data augmentation’ (Tanner and Wong, 1987), as we shall see below.

The model used for the imputations in step 1 is the assumed model for the joint distribution of all observations. Let us assume that the model in which we are interested, often referred to appropriately as the *model of interest*, is a model for a dependent variable Y (e.g., a hierarchical linear model) with an array X of explanatory variables. If the combination of X and Y is the totality of available data, then the model of step 2 is the conditional distribution of Y given the rest of the data according to the joint distribution used in step 1. Then the imputation model and the model of interest are in correspondence. This is, however, not necessary.

Often the research question is such that some observed variables X_{extra} are excluded from the model that is currently being estimated. This may be because of causality arguments – for example, because the variable X_{extra} was measured at a later or earlier moment than the other variables, or because X_{extra} occupies the theoretical place of a mediating variable which is excluded from the currently estimated model. Or a large data set may have been collected, and the dependence of Y on X_{extra} is thought to be weak or theoretically irrelevant, and is therefore not investigated. Also the considerations of Section 9.1.1 can lead to observing variables that are relevant for the missingness mechanism but not for the model of interest. Then missingness might be at random (MAR) in the joint distribution of all the variables (X, X_{extra}, Y), whereas missingness would not be at random (MNAR) if only the observations of (X, Y) would be considered. In all these cases the imputation of missing values should depend not only on the variables in the model of interest, but also on the variables in X_{extra} . Note that this makes the imputation models potentially more complex than the model of interest. To summarize, step 1 will impute by drawing from the conditional distribution of the missing variables given the observations in all of (X, X_{extra}, Y), while step 2 analyzes the dependence of Y only on X .

Of the three steps, we shall now elaborate the first and third. The second step consists of the estimation of parameters of the hierarchical linear model for all randomly imputed data sets, and how to do this should be clear to anyone who has read the relevant chapters of this book. Note that each of the imputed data sets itself is a complete data set so that it can be analyzed by the multilevel methods for complete data.

Several computer packages for multilevel analysis have recently added options for multiple stochastic imputation, or are in the process of doing so. For example, MLwiN has a macro for relatively simple imputations, and an associated program REALCOM-impute for imputation of missing data in more complex patterns; and the mice package in R can impute missing values in two-level models. See Chapter 18 for more information about these packages.

Plausible values

Multiple data sets constructed from multiple stochastic imputation are sometimes called *plausible values*. These are used particularly when the missing values are *missing by design*. This means that the data collection was such that some variables were missing on purpose for some parts of the sample, or for some waves in the case of panel surveys. For example, if there are four related questions that are too much of a burden on the respondent, the

sample could be divided randomly into six groups, each of which gets the same two out of these four questions. The missingness for such variables is MCAR, which facilitates the construction of the imputation model.

9.4.1 The imputation method

Let us now consider step 1, the imputation of the missing data, that is, the process of replacing missing values by likely values. If there are missing values in several variables, this will require the conditional distribution of all these variables. The best way to obtain random draws from the conditional distribution of the missing values given the observed data will depend on the model under consideration. There is one flexible approach, however, which can often be followed. This can be described as iteratively drawing from all the conditional distributions of each variable given the rest, and continuing this process until a stochastic kind of convergence is obtained. It is a Bayesian approach (see Section 12.1), which means that the parameters are themselves also regarded as random variables, and their distribution reflects the uncertainty about their true values.

Denote all variables jointly by X_1, \dots, X_p, Y . This array should also include the variables that were referred to above as X_{extra} . Denote by φ the vector of all parameters of the joint distribution of the complete data (X_1, \dots, X_p, Y) .

1. Initialize the process by a random draw from an approximation, perhaps quite rough, from the conditional distribution of the missing values given the observed values. In many cases a reasonable approach is to approximate the distribution of (X_1, \dots, X_p, Y) by a multivariate normal distribution, ignoring the multilevel structure; and draw the missing values from this estimated normal distribution. This is straightforward because many statistical software programs nowadays contain good procedures for estimating the parameters for the multivariate normal distribution from incomplete data, for example, using the EM algorithm (Little and Rubin, 2002); likewise, for normal distributions many software packages contain methods for drawing from the conditional distributions, which again are normal. If some of the X_k or Y variables have clearly nonnormal distributions, then a transformation may be applied to reduce skewness, and variables may be truncated or rounded to the set of permitted values; since this is the initialization only, a rough procedure is acceptable. This initialization yields a completed data set.
2. Let $\tilde{\varphi}$ be a random draw from the conditional distribution of φ (the so-called *posterior distribution*) given the completed data set.
3. Impute the missing values in Y by random draws from the conditional distribution of the missing cases in Y given X_1, \dots, X_p and given the observed cases in Y , using for (X_1, \dots, X_p) the completed data set and parameter value $\tilde{\varphi}$.
4. Successively for $k = 1, \dots, p$, impute the missing values in X_k by random draws from the conditional distribution of the missing cases in X_k given $X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_p, Y$ and given the observed cases in X_k , using for $(X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_p, Y)$ the completed data set and parameter value $\tilde{\varphi}$.
5. Go to step 2 and repeat the process, until it is deemed that it has lasted long enough.

Of course steps 3 and 4 need to be carried out only for the variables that have missing values.

This procedure defines a stochastic iterative process. The general principle was proposed by Tanner and Wong (1987) who called it *data augmentation*, because the data set is augmented by the imputed missing data. Cycling through the conditional distributions in this way is called Gibbs sampling (e.g., Gelman et al., 2004), which again is a particular kind of Markov chain Monte Carlo (MCMC) procedure (Section 12.1). Mathematical results ensure that the joint probability distribution of $(X_1, \dots, X_p, Y, \tilde{\varphi})$ converges to the joint distribution of the complete data and the parameters given the observed data; recall that this procedure is defined in a Bayesian framework in which the parameters are also random variables. This has the advantage that the uncertainty about the parameter values will be reflected in corresponding greater dispersion in the imputed values than if one followed a frequentist procedure and fixed the parameter at an estimated value.

The cycle consisting of steps 2–4 will be repeated a large number of times. First a ‘burn-in’ of a number of cycles is needed to give confidence that the process has converged and the distribution of generated values is more or less stable. After this point, the generated imputations can be used. Successively generated imputations will be highly correlated. Therefore a *sampling frequency* K has to be chosen, and one in every K generated imputations will be retained for further use. K will have to be large enough that the imputed values separated by K cycles have a negligible correlation.

It may be difficult to assess when convergence has taken place, however – in step 5 this was hidden behind the word ‘deemed’. The literature on MCMC procedures contains material about convergence checks, and an important sign is that the consecutive values generated for the components of the parameter $\tilde{\varphi}$ have achieved some kind of stability, as can be seen from plotting them. More detailed diagnostics for convergence may be used in more complicated models. Abayomi et al. (2008) propose to study the convergence of the imputation procedure by considering the marginal distributions and bivariate scatterplots of the imputed data sets, looking in particular at the deviations between the observed and the imputed data points. Evidently, it is quite acceptable that there are differences; when missingness depends on observed variables, then differences between distributions of observed and imputed data are indeed to be expected. But one should check that the imputed data seem plausible and are in line with presumed mechanisms of missingness.

The procedure of steps 1–5 was proposed and developed by Yucel (2008) for normally distributed variables in two-level and three-level models; and by Goldstein et al. (2009) for quite general two-level models, including models for binary and categorical variables, where level-one as well as level-two variables may have missing values. Such procedures are implemented in REALCOM-impute (Goldstein, 2011) and the R package mlmmm.

Example 9.1 Simulated example: missingness dependent on a third variable.

As a simple artificial example, consider the model

$$Y_{ij} = 0.5 + 0.2X_{1ij} + U_{0j} + R_{ij}, \quad (9.1)$$

where X_{1ij} and R_{ij} are independent standard normal variables, and the random intercept U_{0j} has variance $\tau_0^2 = 0.16$. Let the group sizes be constant at $n = 20$ and let there be $N = 200$ groups;

the number of groups is sufficiently large that one simulated data set should give a good indication of how parameters are being recovered. Suppose that there is a third variable X_{2ij} given by

$$X_{2ij} = X_{1ij} + Y_{ij} + E_{ij},$$

where the E_{ij} likewise are independent standard normal variables; and let missingness be determined according to the rule that Y_{ij} is observed only for $X_{2ij} \leq 2$. We simulated such a data set, and 834 out of the 4,000 values of Y_{ij} turned out to be missing. Suppose the model of interest is

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{1ij} + \gamma_{01}\bar{X}_{1j} + U_{0j} + R_{ij}. \quad (9.2)$$

Since the data are generated according to (9.1), the true parameter values are $\gamma_{00} = 0.5$, $\gamma_{10} = 0.2$, $\gamma_{01} = 0$, $\tau_0^2 = 0.16$, $\sigma^2 = 1$.

The variables for imputation are Y, X_1, X_2 . Thus, in terms of the discussion on p. 136, $X = X_1$ while $X_{\text{extra}} = X_2$. The missingness mechanism for the three variables Y, X_1, X_2 jointly here is MAR. Since missingness of Y depends on X_2 , the model for only Y and X_1 is MNAR.

Model 1 of Table 9.1 gives the result of an analysis of the model of interest after dropping all level-one units with a missing value. The estimated parameter values are far off; for example, γ_{10} is estimated as 0.028 with a standard error of 0.018, whereas the true value is 0.2. This illustrates the pernicious effect of listwise deletion of missing data on parameter estimates.

Table 9.1: Estimates for random intercept model with listwise deletion of missing values (Model 1) and with randomly imputed values according to a multivariate normal distribution (Model 2).

	Model 1		Model 2	
Fixed effect	Coefficient	S.E.	Coefficient	S.E.
γ_{00} = Intercept	0.207	0.029	0.435	0.030
γ_{10} = Coefficient of X_1	0.028	0.018	0.195	0.016
γ_{01} = Coefficient of \bar{X}_1	-0.101	0.127	-0.067	0.130
Random effect	Var.	S.E.	Var.	S.E.
<i>Level-two variance:</i>				
$\tau_0^2 = \text{var}(U_{0j})$	0.112	0.017	0.125	0.018
<i>Level-one variance:</i>				
$\sigma^2 = \text{var}(R_{ij})$	0.829	0.022	0.978	0.022
Deviance	8,629.4		11,526.2	

Model 2, on the other hand, was obtained after imputing the missing values based on an assumed model that the random vectors $(Y_{ij}, X_{1ij}, X_{2ij})$ are independent for different values of (i, j) , and have a three-variate normal distribution. This model is correct for the marginal distribution of $(Y_{ij}, X_{1ij}, X_{2ij})$ but ignores the multilevel structure. It is the model suggested as step 1, the initialization, of the procedure above. This model recovers the fixed parameters quite well, with, for example, $\hat{\gamma}_{10} = 0.195$. The recovery of the variance parameters is less good, which may be expected since the dependence structure is not well represented by the imputation model. The results are not trustworthy, in addition,

because the estimation is done as if the imputed data were observed in a bona fide way, whereas they were made up. This has the consequence that the standard errors will be too low, suggesting undue precision. This issue will be dealt with in the next subsection, where the example is continued.

Model 1 is based on considerably fewer cases than Model 2. The deviance in Model 2 is based on an inflated data set, and therefore is totally incomparable with that of Model 1.

9.4.2 Putting together the multiple results

We now continue with steps 2 and 3 of page 135, the combination of the multiple imputations so as to produce estimates for the model of interest.

We suppose that we have a number of data sets in which the missing values have been randomly imputed according to the methods of the preceding section. It is also assumed that, given the observed parts of the data, the imputed parts are independent across these data sets. In the MCMC method of the preceding section, this means that the sampling frequency K must be high enough, ensuring negligible correlation between consecutively generated imputed data sets. Denote the number of imputed data sets by M and the data sets themselves by $D_{(1)}, D_{(2)}, \dots, D_{(M)}$. Thus, each data set contains the variables X and Y , and all have the same values for the observed data points but different imputed values for the missing data points. We first discuss how to combine parameter estimates across the imputed data sets, and then how to combine deviance tests across the imputations.

In step 2, for each data set $D_{(m)}$ the parameters are estimated by the method appropriate to the model of interest, for example, the hierarchical linear model. Let us focus on a single parameter, for example, one of the regression coefficients (fixed effects) or variance parameters. This parameter is arbitrarily denoted θ . Denote the estimate for this parameter from data set $D_{(m)}$ by $\hat{\theta}_{(m)}$ and the associated standard error, calculated under the erroneous assumption of complete data, by $S.E._{(m)}$. If there is very little missing data, or the other variables permit quite accurate imputations (e.g., because for the variable that has missing values there exists another, highly correlated variable, which was always observed), then the estimates $\hat{\theta}_{(m)}$ will be scarcely different from each other. On the other hand, if the missingness entails a considerable loss of information then the estimates $\hat{\theta}_{(m)}$ will be very different from one another, depending on the random imputations. Thus, there are two sources of imprecision in the parameter estimates: the *within-data-set uncertainty* which is reflected by the standard errors $S.E._{(m)}$, and the *imputation uncertainty* reflected by the variation between the estimates $\hat{\theta}_{(m)}$. These two sources are combined by the following formulas from Little and Rubin (2002).

First, from the multiple imputations we compute the average estimate

$$\bar{\theta} = \frac{1}{M} \sum_{m=1}^M \hat{\theta}_{(m)}, \quad (9.3)$$

the average within-data-set variance

$$\overline{W} = \frac{1}{M} \sum_{m=1}^M S.E._{(m)}^2, \quad (9.4)$$

and the between-imputation variance

$$B = \frac{1}{M-1} \sum_{m=1}^M (\hat{\theta}_{(m)} - \bar{\theta})^2. \quad (9.5)$$

Then the combined estimate is the mean parameter estimate (9.3) over all imputed data sets, and its standard error is

$$\text{S.E.}(\bar{\theta}) = \sqrt{\bar{W} + \left(1 + \frac{1}{M}\right)B}, \quad (9.6)$$

If the different imputations lead to almost the same estimates $\hat{\theta}_{(m)}$, then the between-imputation variance B will be very small, and the standard error (9.6) is practically the same as the individual standard errors $\text{S.E.}_{(m)}$. However, often the situation is different, and the overall standard error (9.6) will be larger.

To test the null hypothesis that the parameter θ is equal to some value θ_0 , the usual t ratio,

$$t = \frac{\bar{\theta} - \theta_0}{\text{S.E.}(\bar{\theta})}, \quad (9.7a)$$

can be tested against a t distribution with degrees of freedom given by

$$df = (M-1) \left(1 + \frac{W}{(1+(1/M))B}\right)^2. \quad (9.7b)$$

This is based on the assumption that the $\hat{\theta}_{(m)}$ estimates are approximately normal, discussed below (p. 144). Similar multi-dimensional formulas are possible for vector-valued parameters θ ; see Little and Rubin (2002).

A useful measure is the *fraction of missing information* for estimating a given parameter. Intuitively, it is evident that if W is larger, as compared to B , there is a larger difference between the ‘complete data sets that could have been’, so that a larger value of W compared to B points to a more important amount of missing information. The formula for the estimated fraction of missing information is

$$\text{Missing fraction} = \frac{(1+(1/M))B}{\bar{W} + (1+(1/M))B}. \quad (9.8)$$

This expresses how much of the information about the value of a given parameter, potentially available in the complete data set, has been lost because of the missingness. This fraction will depend on the parameter under consideration, as we shall see in the example below. The estimated value may be rather unstable across independent repetitions of the imputation process, but will indicate the rough order of magnitude.

Wald tests as well as deviance (likelihood-ratio) tests can be combined by methods discussed by Little and Rubin (2002). Suppose that a null hypothesis is being tested by a Wald test or deviance test and the multiple imputations yield, for the completed data sets obtained, test statistics C_1, C_2, \dots, C_M which in the complete data case have under the null

hypothesis an asymptotic chi-squared distribution with q degrees of freedom. Test statistic C_m will be either the Wald test, or the difference between the deviances for the estimated models corresponding to the null and alternative hypotheses, each for the same data set $D_{(m)}$. Define by \bar{C} the average test statistic,

$$\bar{C} = \frac{1}{M} \sum_{m=1}^M C_m, \quad (9.9)$$

and by V the sample variance of the square roots $\sqrt{C_m}$ of C_m ,

$$V = \frac{1}{M-1} \sum_{m=1}^M (\sqrt{C_m} - \bar{\sqrt{C}})^2, \quad (9.10)$$

where

$$\bar{\sqrt{C}} = \frac{1}{M} \sum_{m=1}^M \sqrt{C_m}.$$

Then the combined test statistic is

$$\tilde{C} = \frac{M\bar{C}/q - (M-1)V}{M + (M+1)V}. \quad (9.11)$$

This test statistic must be tested against an F distribution with degrees of freedom q and b , where

$$b = q^{-3/M}(M-1) \left(1 + \frac{M}{(M+1)V} \right)^2. \quad (9.12)$$

These rather intimidating equations work out in such a way that when the deviance statistics C_m are practically the same, so that V is very small, then $q \times \tilde{C}$ is very close to each of these deviances and the result of the combined tests is practically the same as the result based on the individual deviance tests. For deviance tests, Little and Rubin (2002) also present another combined test, but this requires more information than the individual test statistics.

The original work in which Rubin (e.g., 1987) developed the multiple imputation method suggested that a very low number of imputations such as $K = 5$ is sufficient. However, this may lead to instability of the estimated between-imputation variance B . Further, a larger fraction of missing data will require a larger number of imputations. Graham et al. (2007) developed recommendations for the number of imputations; for example, their advice is to use at least 20 imputations if 10–30% of the information is missing, and at least 40 imputations if half of the information is missing. These proportions of missing information refer to the quantity estimated by (9.8).

If multiparameter tests are required, for example, for testing the effect of a categorical variable with three or more categories, then the required number of imputations is also

larger. Carpenter and Kenward (2008) recommend multiplying the number of imputations by $q(q + 1)/2$, where q is the dimension of the multiparameter test.

Example 9.2 Simulated example continued: multiple imputations.

We continue the preceding example. Recall that the model of interest is

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{1ij} + \gamma_{01}\bar{X}_{1,j} + U_{0j} + R_{ij}$$

(see (9.2)), there are missing values only in Y , and the missingness mechanism is known to be MAR, given the additional variable X_2 .

The first set of results in Table 9.2 repeats Model 2 of Table 9.1, based on a single random imputation under a multivariate normal model for (Y, X_1, X_2) without a multilevel structure, and treating this as an observed data set. The second set of results present the result of multiple imputations, using the imputation model

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{1ij} + \gamma_{01}\bar{X}_{1,j} + \gamma_{20}X_{2ij} + \gamma_{02}\bar{X}_{2,j} + U_{0j} + R_{ij},$$

which is in accordance with the conditional distribution of the missing values given the observed data, although some of the coefficients are 0. We used $M = 50$ imputed data sets.

Table 9.2: Estimates for random intercept model with a single imputation and with $M = 50$ multiple imputations.

Fixed effect	True value	Model 2		Model 3			
		Single imp.	Multiple imp.	Coeff.	S.E.	Coeff.	S.E.
γ_{00} = Intercept	0.5	0.435	0.030	0.444	0.033	0.07	
γ_{10} = Coeff. of X_1	0.2	0.195	0.016	0.199	0.019	0.30	
γ_{01} = Coeff. of \bar{X}_1	0.0	-0.067	0.130	-0.113	0.141	0.05	
Random effect		Var.	S.E.	Var.	S.E.	Miss.fract.	
<i>Level-two variance:</i>							
$\tau_0^2 = \text{var}(U_{0j})$	0.16	0.125	0.018	0.150	0.021	0.08	
<i>Level-one variance:</i>							
$\sigma^2 = \text{var}(R_{ij})$	1.0	0.978	0.022	0.966	0.027	0.32	

Comparing the two sets of results shows that in this case the parameter estimates from the single imputation were quite close already, and the more realistic standard errors for the multiple imputations all are somewhat larger than those for the single imputation, in accordance with what is expected from (9.6), but also for the standard errors the differences are not large. For the level-two variance τ_0^2 a clear improvement is obtained, the true value being 0.16. The imputations using an imputation model with a multilevel structure led to good recovery of the intraclass correlation, which was underestimated by imputations using a multivariate normal model without multilevel structure.

The fractions of missing information range from 0.05 for the regression coefficient of \bar{X}_1 to 0.32 for the level-one variance. The proportion of cases with a missing value was $834/4,000 = 0.21$. The fractions of missing information for the parameters pertaining to the level-two model are smaller

than this value, which means that the multiple imputation succeeds in recovering a greater part of the information for these parameters; this is possible because there the other level-one units in the same level-two unit also contribute to the information. That the fraction of missing information for the level-one parameters are larger than 0.21 might be a random deviation.

This example is quite extreme, because the missingness is determined deterministically by the observed variable X_2 which is strongly correlated with the dependent variable. Still, this example teaches us a number of things:

1. An analysis using only the complete data and ignoring the existence of missing data can be hopelessly biased.
2. If missingness is at random (MAR) and a correct imputation model is used, then multiple stochastic imputation can yield approximately unbiased results.
3. It is not necessary to know the missingness mechanism, it is sufficient to know that MAR holds for the given set of variables, and to have an imputation model that correctly represents the dependencies between these variables.
4. If there are missing values only in the dependent variable, but other variables are available that are predictive of the dependent variable and not included in the model of interest, cases with a missing dependent variable should not be dropped and multiple imputation can lead to more credible results.
5. Imputation models that do not represent the dependencies in the data well may yield results that in a similar way underestimate these dependencies.

Normality assumptions for multiple imputation

This method of combining the multiple imputations assumes that the parameter estimates $\hat{\theta}$ in the complete data set have a normal sampling distribution. In multilevel models, this will always require that the number of units at the highest level is large enough, say, 30 or more (with values as low as 20 still possible but giving poorer approximations); and the total number of level-one units is high enough, say, at least 100. For the estimated regression coefficients as well as for the level-one residual variance, normality will then be a reasonable assumption. For the estimated variance components at level two and higher, however, the approximate normality will break down if the true variance is relatively small. It is safer to apply equations (9.3)–(9.6) to the estimated standard deviation, and transform back to the variance, if this is desired. Note that in Section 6.3 we commented on the normal approximation for the distribution of variance parameters, and presented the approximate equation (6.2) for transforming standard errors between variances and standard deviations, with conditions under which this is a reasonable approximation. If these conditions are not met, a Bayesian approach (Section 12.1) will still be possible.

9.5 Multiple imputations by chained equations

If there are missing values in more than one variable, then step 4 in the multiple imputation procedure as described on p. 137 can be difficult in practice. It requires the conditional

distributions to be specified for all variables that have any missing values, conditional on all the other variables; and these conditional distributions must be coherent in the sense that they belong to the simultaneous distribution of all variables together. The multilevel dependencies between the variables may make this difficult. Goldstein et al. (2009) developed a quite general model for this, applicable to combinations of continuous and discrete variables, which is implemented in REALCOM-impute (Goldstein, 2011). However, difficulties still may remain for various model specifications, for example, if interactions are involved or there are differences between within-group and between-group regressions.

A very flexible approximate procedure for imputing missing values is *imputation by chained equations*. This has been proposed by various authors and is also known under various other names, such as *fully conditional specification* and *partially incompatible MCMC*. Van Buuren (2007) gives an overview of this literature. Although the theoretical grounding of the method is incomplete, it is intuitively very appealing and has yielded good results in simulations, as shown by Raghunathan et al. (2001), van Buuren et al. (2006), van Buuren (2007), and for clustered data by Zhao and Yucel (2009). This method was further developed for multilevel data structures by van Buuren (2011).

It proceeds as follows. We consider a statistical model with basic variables which for the moment we shall call Y_1, Y_2, \dots, Y_K , all of which may contain some missing data points. We do not specify the simultaneous distribution of all variables jointly, but only model the conditional distribution of each variable on all the others. Each of these distributions is a regression-type model with a response variable Y_k and a set of explanatory variables $Y_1, \dots, Y_{k-1}, Y_{k+1}, \dots, Y_K$, which may be augmented by interactions, group means, and other variables constructed from these. Depending on the type of variable, this might be a linear regression model if Y_k is continuous, a logistic regression model if Y_k is binary, etc. For the multilevel case, these models would be the multilevel variants: a hierarchical linear model, a multilevel logistic regression model, etc. Denote by φ_k the vector of parameters for this model. The principle of imputation by chained equations is to treat in turn the imputation for each of the variables Y_k while considering the others as given, using these regression-type models; and cycle repeatedly through the variables. This can be described as follows:

1. Initialize the process in a reasonable way. This can be done, for example, by doing the following for each k . Estimate φ_k from the available data, and impute by drawing from the conditional distribution of the missing values in Y_k given the observed data. This initialization yields a completed data set.
2. For each $k = 1, \dots, K$ in turn, do the following.
 - (a) Make a random draw $\tilde{\varphi}_k$ from the conditional ('posterior') distribution of φ_k given the completed data set.
 - (b) Impute the missing values in Y_k by random draws from the model for Y_k given the other variables $Y_1, \dots, Y_{k-1}, Y_{k+1}, \dots, Y_K$, using the completed data set for these variables, conditioning on the observed cases for Y_k , and using the estimated parameter value $\tilde{\varphi}_k$.
 - (c) If Y_k is a variable in a multilevel structure, and group means of Y_k for groups defining a higher level are among the explanatory variables for some of the other variables Y_h , then update these group means from the completed data for

Y_k . (The same will be done if interactions or other kinds of aggregates than group means play a role.)

3. Repeat the process of step 2 until it is deemed that it has lasted long enough.

If for some variables there are no missing data, they can evidently be skipped in the imputation process.

The secret here is that we do not worry whether the models for Y_k given $Y_1, \dots, Y_{k-1}, Y_{k+1}, \dots, Y_K$ hang together in one multivariate distribution. This is a theoretical defect but a practical benefit. For example, for a two-level model of interest, the variables Y_1, \dots, Y_K would be the dependent variables and all the explanatory variables with any missing values, the level-one as well as the level-two variables. Each of these can be modeled separately in the most appropriate way. The level-one variables can be modeled by the hierarchical linear model if they are continuous, by a multilevel logistic regression model if they are binary, etc.; level-two variables can be modeled by single-level linear or logistic or other generalized linear regression models, with as predictors the other level-two variables and the group means of the level-one variables; interactions can be accommodated without any problems.

Multilevel software is now beginning to implement the procedures of this and the preceding sections. Since their implementation is not yet widespread, we give some more details on how these steps can be carried out.

Drawing from the posterior distribution of the parameters as mentioned in step 2(a) is a basic element in Bayesian estimation procedures treated in Section 12.1 (see p. 196).

This Bayesian step can also be replaced by the following.

- a'. Estimate the parameter $\hat{\varphi}_k$ from the completed data set, together with its estimated covariance matrix Σ_k . Draw $\tilde{\varphi}_k$ from the multivariate normal distribution with mean $\hat{\varphi}_k$ and covariance matrix Σ_k .

For regression coefficients especially this will be hardly any different from the Bayesian procedure (step 2(a)). An even simpler method mentioned by van Buuren (2007) is the plug-in method:

- a''. Estimate the parameter $\hat{\varphi}_k$ from the completed data set, and use $\tilde{\varphi}_k = \hat{\varphi}_k$.

If sample sizes are reasonably large, this will usually yield quite similar results to the full procedure.

The procedure for imputing as in step 2(b) depends on the type of variable that is being considered. Let us assume that we are dealing with a two-level data structure where the group sizes are rather homogeneous. (Strongly heterogeneous group sizes may lead to variance heterogeneity between the groups, which leads to further complications.) First consider the case where Y_k is a level-two variable for which a linear regression model is reasonable. The parameter value $\tilde{\varphi}_k$ comprises the regression coefficients and the residual variance, to be denoted $\tilde{\sigma}_k^2$. Then for each level-two unit j , using the regression coefficients and the explanatory variables (which are completely known in the imputed data set), calculate the predicted value \hat{Y}_{kj} . Generate independent normally distributed random variables R_{kj} with expected value 0 and variance $\tilde{\sigma}_k^2$. Then the imputed value is $\hat{Y}_{kj} + R_{kj}$.

Now consider the case where Y_k is a level-one variable for which a random intercept model is used; denote this by

$$Y_{kij} = \text{fixed part} + U_{0j} + R_{ij}.$$

Here the parameter value $\tilde{\varphi}_k$ comprises the regression coefficients, the intercept variance $\tilde{\sigma}_0^2$, and the residual variance $\tilde{\sigma}_k^2$. A complication arises from the fact that the observed values for Y_{kij} for $i' = 1, \dots, n_j$ offer information about U_{0j} which should be used in the imputation. We proceed by first calculating the predicted value \hat{Y}_{kij} from the regression coefficients and the explanatory variables. Then we calculate the empirical Bayes estimate (posterior mean; see Section 4.8) denoted by \hat{U}_{0j}^{EB} , and the associated posterior standard deviation (or comparative standard deviation, see p. 65), denoted $S.E._j$. These values depend both on the parameters $\tilde{\varphi}_k$ and on the observed (nonmissing) data for variable Y_k . Next we generate a normally distributed random variable ΔU_{0j} with mean 0 and variance $S.E.^2_j$. Then the imputed value for U_{0j} is the sum⁴ $\tilde{U}_{0j} = \hat{U}_{0j}^{\text{EB}} + \Delta U_{0j}$. Finally, for all missing values of Y_{kij} we generate independent normally distributed random variables R_{kij} with expected value 0 and variance $\tilde{\sigma}_k^2$. The missing values are then imputed by

$$\tilde{Y}_{kij} = \hat{Y}_{kij} + \hat{U}_{0j}^{\text{EB}} + \Delta U_{0j} + R_{kij}. \quad (9.13)$$

Example 9.3 Comparison of results without and with imputations.

Here we again use the data set for which many examples were given in Chapters 4 and 5. The model of interest is a random intercept model with effects of verbal IQ and SES at the individual level as well the group level,

$$\begin{aligned} Y_{ij} = & \gamma_{00} + \gamma_{10} \text{IQ}_{ij} + \gamma_{20} \text{SES}_{ij} + \gamma_{30} \text{IQ}_{ij} \times \text{SES}_{ij} \\ & + \gamma_{01} \bar{\text{IQ}}_j + \gamma_{02} \bar{\text{SES}}_j + \gamma_{03} \bar{\text{IQ}}_j \times \bar{\text{SES}}_j + U_{0j} + R_{ij}. \end{aligned}$$

As auxiliary variables X_{extra} in the imputation model we used three variables that are closely related to the variables in the model of interest: an earlier language test, a measure of ‘performat IQ’, and a dummy variable indicating belonging to an ethnic minority. In these variables we have the following numbers of missing values:

Language test Y	114
IQ	17
SES	133
Performat IQ	8
Earlier language test	318
Minority	0

For the three variables with most missing cases, there is very little overlap in the individuals for whom there are missing values, which leads to good possibilities for imputation.

The imputation was done by the method of chained equations, where for each variable with missing values the imputation model was defined as a random intercept model using the other five variables, as well as their group means, as predictors. From a preliminary estimation it appeared that several interactions seemed to have significant effects: the product interactions between IQ and SES, between the group means of IQ and of SES, and between IQ and the classroom proportion (i.e., group mean) of minority students. These interactions were added to the imputation models for the variables that were not involved in these interactions.

Table 9.3 presents two sets of results for the language test, as a function of SES and IQ, in a random intercept model. The left-hand column gives the parameter estimates of the naive analysis, where all students with a missing value on at least one of these variables have been dropped. The

⁴Incidentally, it may be noted that \hat{U}_{0j}^{EB} and ΔU_{0j} are uncorrelated (cf. (4.18)), so that $\text{var}(U_{0j}) = \text{var}(\hat{U}_{0j}^{\text{EB}}) + S.E.^2_j$.

right-hand column gives the estimates from the analysis based on imputation by chained equations, with 50 imputations. The results for the two analyses are very similar, which suggests that in this case the missingness is not related to the ways in which intelligence, socio-economic status, and schools influence this language test. This permitted us to use casewise deletion in the examples in Chapters 4-5.

Table 9.3: Estimates for random intercept model with complete cases only, and using multiple imputations.

Fixed effect	Complete cases		Multiple imputations		
	Coeff.	S.E.	Coeff.	S.E.	Miss.fract.
γ_{00} = Intercept	41.530	0.244	41.450	0.238	0.02
γ_{10} = Coeff. of IQ	2.209	0.056	2.223	0.055	0.03
γ_{20} = Coeff. of SES	0.174	0.012	0.173	0.012	0.05
γ_{30} = Coeff. of IQ \times SES	-0.018	0.005	-0.016	0.005	0.08
γ_{01} = Coeff. of \overline{IQ}	0.962	0.309	0.901	0.308	0.02
γ_{02} = Coeff. of \overline{SES}	-0.096	0.044	-0.104	0.044	0.02
γ_{03} = Coeff. of $\overline{IQ} \times \overline{SES}$	-0.091	0.032	-0.070	0.030	0.07
Random effect	S.D.	S.E.	S.D.	S.E.	Miss.fract
<i>Level-two standard deviation:</i>					
τ_0^2 = S.D.(U_{0j})	2.841	0.181	2.767	0.178	0.01
<i>Level-one standard deviation:</i>					
σ^2 = S.D.(R_{ij})	6.162	0.073	6.200	0.073	0.05

9.6 Choice of the imputation model

In any imputation procedure, the choice of the imputation model is a crucial decision. The discussion here applies to imputation based on the simultaneous distribution as well as to imputation by chained equations.

In the first place, an ample set of variables should be used for the imputations. If reasonable predictors are available for variables with missing cases, or for missingness itself, then it is wise to include those in the imputation model; if they are not part of the model of interest they will have the role of the X_{extra} variable discussed on p. 136. In the case of complex sample surveys, variables that determine inclusion probabilities should also be included in the imputation model (cf. Chapter 14). All variables that make the MAR assumption more plausible as compared to MNAR should be included in this way, even if they have some missing values themselves.

Second, a general rule is that it does not hurt to use relatively complex models for the imputations, as long as they are feasible in practice. Issues of overfitting as are known for parameter estimation in most cases do not matter for imputation. One exception is the phenomenon of ‘complete separation’ in logistic regression: sometimes for a logistic

regression model there is a threshold value such that for values of the linear predictor higher or lower than the threshold, the observed values of the dependent variable are all zeros or all ones. When this is encountered in a model used for imputation, extra care is required to check if this would be realistic, and avoid it if it is not.

Third, while simplification certainly is allowed and even unavoidable, it should be realized that dependencies in the total data set that are not represented in the imputation model will also be underrepresented in the imputed values, which may lead to bias in the estimates for the model of interest. For example, as was illustrated by Example 9.2, using a single-level imputation model for a two-level data set is likely to lead to misrepresentation of the intraclass correlations. As another example, not using group means of level-one variables as explanatory variables for the imputation of other level-one variables may lead to biased estimation of differences between within-group and between-group regression coefficients. This implies that especially those aspects in which the researcher is mainly interested should be represented well in the imputation model.

When constructing the imputation model from separate models for each variable that has any missing cases (e.g., if one uses chained equations), a useful procedure may be the following. Start by fitting preliminary models for the distribution of all variables given the others, and for the distribution of the missingness indicators given the others. These preliminary models could be based on one simple and rough imputed data set (e.g., obtained using the assumption of multivariate normality and without taking the multilevel structure into account). For each pair of variables Y_k and Y_h , if either variable is important for predicting the other variable, and/or variable Y_h is important for predicting missingness in Y_k , then include Y_h in the imputation model for Y_k .

Van Buuren (2011) suggests using imputation models that are multilevel models with heterogeneous level-one variances. This may be relevant mainly when group sizes are strongly heterogeneous, because then a regular hierarchical linear model of Y on a variable X which itself has a homogeneous within-group variance may lead to a *conditional* distribution of X given Y with heterogeneous level-one variances. This is unlikely to be much of a concern when group sizes do not vary strongly.

Finally, researchers should not be afraid of having to make some approximations and reasonable guesses. Always they must realize that imputation is guesswork unless the missing data mechanism is known *a priori*, for example, because of having been built into the data collection design. Results of statistical analyses based on imputation of missing values, in cases where the missingness mechanism is *a priori* unknown, will be surrounded by even more uncertainty than are statistical results obtained from complete data sets.

9.7 Glommary

Missingness indicators. Binary random variables indicating that a data point is missing or observed.

Complete data. The hypothetical data structure that is regarded as the data set that would be observed if there were no missing data points.

Missingness completely at random (MCAR). The missingness indicators are independent of the complete data.

Missingness at random (MAR). Conditionally given the observed data, the missingness indicators are independent of the unobserved data. Under MAR, naive approaches such as analysing the complete cases only have the risk of serious bias and loss of information, but it is possible to follow approaches that do not have these defects.

Missingness not at random (MNAR). This is the negation of MAR. The analysis of incomplete data where missingness is not at random will always depend on untestable assumptions, and will be more complicated and leave open more questions than the MAR case. This book does not treat methods of analysis for the MNAR case. For further literature the reader is referred to Molenberghs and Kenward (2007), Carpenter and Kenward (2008), Graham (2009), Hogan et al. (2004), Yuan and Little (2007), Roy and Lin (2005), and Mason et al. (2010).

Implications for design. It is recommended to collect auxiliary data that are predictive of missingness indicators and of the values of unobserved data points. Including such auxiliary data can push the design in the direction of MAR.

Model of interest. The model on which the research is substantively focused.

Missing values on the dependent variable. If there are missing data points only for the dependent variable, while there are auxiliary variables that are predictive of missingness or of the unobserved values, and these auxiliary variables are not included in the model of interest, then bias may be reduced by methods of analysis that use these auxiliary variables and retain the data points with a missing dependent variable.

Full information maximum likelihood. When possible, an analysis based on the likelihood of the observed data is preferable. This is called full information maximum likelihood. This approach is sometimes possible, but cumbersome, in cases with missing data points for several variables in a nonmonotone pattern.

Imputation. Filling in the missing data.

Multiple stochastic imputation. Constructing multiple imputed data sets by drawing the imputations for the missing data points from the conditional distribution of their values, given the observed data.

Imputation model. The probability model used for the stochastic imputation. It is recommended, if possible, to use auxiliary variables that are not included in the model of interest but that are predictive of missingness or of the unobserved values, and to use these auxiliary variables in the imputation model.

Combination of results. Each imputed data set is a complete data set for which parameter estimates can be obtained for the model of interest. These estimates can be combined across the multiple imputed data sets to obtain, in the MAR case, and if the imputation model is correct, approximately unbiased parameter estimates and standard errors.

Multiple imputation by chained equations. An imputation method that needs, for each variable that has missing data, a model for this variable conditional on the other variables; but does not require these models to be the conditional distributions for

the simultaneous distribution of all variables jointly. This is a flexible approach that can be utilized fruitfully for the complexities of multilevel data structures.

Choice of the imputation model. It is advisable to take care that the important dependencies in the data are represented in the imputation model; for example, the multilevel data structure, interactions, and differences between within-group and between-group regressions.

10

Assumptions of the Hierarchical Linear Model

Like all statistical models, the hierarchical linear model is based on a number of assumptions. If these assumptions are not satisfied, the procedures for estimating and testing coefficients may be invalid. The assumptions are about the linear dependence of the dependent variable, Y , on the explanatory variables and the random effects; the independence of the residuals, at level one as well as at the higher level or levels; the specification of the variables having random slopes (which implies a certain variance and correlation structure for the observations); and the homoscedastic normal distributions for the residuals. It is advisable, when analyzing multilevel data, to devote some energy to checks of the assumptions. This chapter is concerned with such checks.

Why are model checks important at all? One danger of model misspecification is the general misrepresentation of the relations in the data (e.g., if the effect of X on Y is curvilinear increasing for X below average and decreasing again for X above average, but only the linear effect of X is tested, then one might obtain a nonsignificant effect and conclude mistakenly that X has no effect on Y). Another is the invalidity of hypothesis tests (e.g., if the random part of the model is grossly misspecified, the standard errors and therefore the hypothesis tests for fixed effects may be completely off the mark). To construct an adequate model for a given data set, good insight into the phenomenon under study is necessary, combined with an inspired application of the relevant social theories; in addition, statistical tools such as those described in this chapter can be helpful.

OVERVIEW OF THE CHAPTER

First, the basic assumptions of the hierarchical linear model are summarized. This model in itself is already much more flexible than the general linear model that is the basis of standard regression analysis. The chapter continues by discussing several approaches that can be straightforwardly applied within the regular hierarchical linear model to check model fit and perhaps improve it by adding additional hierarchical linear model components. This extends what has already been said about this topic in Section 6.4. Specification of the fixed part is treated, with attention to transforming explanatory variables. Then the specification of the random part is discussed, with special attention given to testing heteroscedasticity.

The second half of the chapter is about residuals, which can be defined for each of the levels. Inspection of the residuals may be important in diagnosing units that do not conform well to the model. Measures are explained, analogous to Cook's distance in linear regression, that indicate the influence of higher-level units on the results of the multilevel fit. Finally, a short section is devoted to methods developed to handle data for which the residuals have nonnormal distributions.

One of the threads running through this chapter is, in line with Section 6.4, that model specification in most cases should be done upward: first try to get the model at the lower levels well specified, then continue with the higher levels.

The topics of this chapter are treated with more of the mathematical background in Snijders and Berkhof (2008). With respect to software implementation, there is some variation among multilevel analysis programs in methods for model checking. MLwiN macros to compute the diagnostics mentioned in Sections 10.4–10.7, and R scripts for doing the same, can be downloaded from the website for this book, <http://www.stats.ox.ac.uk/~snijders/mlbook.htm>.

10.1 Assumptions of the hierarchical linear model

The hierarchical linear model was introduced in Chapter 5. In this chapter we consider only the two-level model. The basic definition of the model for observation Y_{ij} on level-one unit i within level-two unit j , for $i = 1, \dots, n_j$, is given in formula (5.15) as

$$Y_{ij} = \gamma_0 + \sum_{h=1}^r \gamma_h x_{hij} + U_{0j} + \sum_{h=1}^p U_{hj} x_{hij} + R_{ij}. \quad (10.1)$$

As before, we shall refer to level-two units also as ‘groups’. The assumptions were formulated as follows after equation (5.12). The vectors $(U_{0j}, U_{1j}, \dots, U_{pj})$ of level-two random coefficients, or level-two residuals, are independent between groups. These level-two random coefficients are independent of the level-one residuals R_{ij} , which also are mutually independent, and all residuals have population mean 0, given the values of all explanatory variables. Further, the level-one residuals R_{ij} have a normal distribution with constant variance σ^2 . The level-two random effects $(U_{0j}, U_{1j}, \dots, U_{pj})$ have a multivariate normal distribution with a constant covariance matrix. The property of constant variance is also called homoscedasticity, nonconstant variance being referred to as heteroscedasticity.

To check these assumptions, the following questions may be posed:

1. Does the fixed part contain the right variables (now X_1, \dots, X_r)?
2. Does the random part contain the right variables (now X_1, \dots, X_p)?
3. Are the level-one residuals normally distributed?
4. Do the level-one residuals have constant variance?
5. Are the level-two random coefficients normally distributed?
6. Do the level-two random coefficients have a constant covariance matrix?

These questions are answered in this chapter in various ways. The answers are necessarily incomplete, because it is impossible to give a completely convincing argument that a given specification is correct. For complicated models it may be sensible, if there are enough data, to employ cross-validation (e.g., Mosteller and Tukey, 1977). Cross-validation means that the data are split into two independent halves, one half being used for the search for a satisfactory model specification and the other half for testing of effects. This has the advantage that testing and model specification are separated, so that tests do not lose their validity because of capitalization on chance. For a two-level model, two independent halves are obtained by randomly distributing the level-two units into two subsets.

10.2 Following the logic of the hierarchical linear model

The hierarchical linear model itself is an extension of the linear regression model and relaxes one of the crucial assumptions of that model, the independence of the residuals. The following model checks follow immediately from the logic of the hierarchical linear model, and were mentioned accordingly in Section 6.4.

10.2.1 Include contextual effects

In the spirit of Chapters 4 and 5, for every level-one explanatory variable one should consider the possibility that the within-group regression is different from the between-group regression (see Section 4.6), and the possibility that X has a random slope. The difference between the within- and between-group regression coefficients of a variable X is modeled by also including the group mean of X in the fixed part of the model. This group mean is a meaningful contextual variable, as follows from Chapters 2 and 3.

For interaction variables it can also be checked whether there are contextual effects. This is done by applying the principles of Section 4.6 to, for example, the product variable $X_{ij}Z_j$ rather than the single variable X_{ij} . The cross-level interaction variable $X_{ij}Z_j$ can be split into

$$X_{ij}Z_j = (X_{ij} - \bar{X}_j)Z_j + \bar{X}_jZ_j,$$

and it is possible that the two variables $(X_{ij} - \bar{X}_j)Z_j$ and \bar{X}_jZ_j have different regression coefficients. This is tested by checking whether the level-two interaction variable \bar{X}_jZ_j has an additional fixed effect when $X_{ij}Z_j$ already belongs to the fixed part of the model. If there is such an additional effect, one has the choice between including in the model either $(X_{ij} - \bar{X}_j)Z_j$ and \bar{X}_jZ_j , or $X_{ij}Z_j$ and \bar{X}_jZ_j ; these two options will yield the same fit (cf. Section 4.6).

The assumption that the population mean of U_{hj} is 0, conditionally given all the explanatory variables, implies that these random intercepts and slopes are uncorrelated with all explanatory variables. If this assumption is incorrect, the problem can be remedied by adding relevant explanatory variables to the model. A nonzero correlation between U_{hj} and a level-two variable Z_j is remedied by including the product variable $X_{hj}Z_j$, as discussed in

the preceding paragraph (cf. Snijders and Berkhof, 2008, Section 3.2). A nonzero correlation between U_{hj} and a level-one variable X_{kij} is remedied by including the product variable $X_{hij}\bar{X}_{k,j}$. Since for $h = 0$ we have $X_{hij} = 1$, applying this procedure to the random intercept U_{0j} leads to including the main effect variables Z_j and $\bar{X}_{k,j}$, respectively.

Sometimes it makes sense also to include other contextual variables, for example, the standard deviation within group j of a relevant level-one variable. For example, heterogeneity of a school class, as measured by the standard deviation of the pupils' intelligence or of their prior achievements, may be a complicating factor in the teaching process and thus have a negative effect on the pupils' achievements.

10.2.2 Check whether variables have random effects

Definition of levels

A first issue is the definition of the 'levels' in the model. Formulated more technically, these are the systems of categories where the categories have random coefficients: the random intercepts and possibly also random slopes. Since this is at the heart of multilevel modeling it has already been given much attention (see Chapter 2 and Section 4.3; Section 6.4 is also relevant here). Several papers have studied the errors that may be made when a level is erroneously left out of the model; see Tranmer and Steel (2001), Moerbeek (2004), Berkhof and Kampen (2004), van den Noortgate et al. (2005), and Dorman (2008). A general conclusion is that if a level is left out of the model, that is, a random intercept is omitted for some system of categories, then the variance associated with that level will be redistributed mainly to the next lower and (if it exists) next higher levels. Furthermore, erroneous standard errors may be obtained for coefficients of variables that are defined on this level (i.e., are functions of this system of categories), and hence tests of such variables will be unreliable. This will also hold for variables with strong intraclass correlations for the omitted level. Similarly, if a random slope is omitted, then the standard errors of the corresponding cross-level interaction effects may be incorrectly estimated.

This leads to the general rule that if a researcher is interested in a fixed effect for a variable defined at a certain level (where a level is understood as a system of categories, such as schools or neighborhoods), then it is advisable to include this level with a randomly varying intercept in the multilevel model – unless there is evidence that the associated random intercept variance is negligible. Similarly, if there is interest in a fixed effect for a cross-level interaction $X \times Z$ where Z is defined at a certain level, then X should have a random effect at this level – again, unless the random slope variance is negligible. Furthermore, if there is interest in the amount of variability (random intercept variance) associated with a given level, and there exist next higher and next lower levels of nesting in the phenomenon under study, then these levels should be included in the model with random intercepts.

Specification of random slopes

It is possible that the effect of explanatory variables differs from group to group, that is, some variables have random slopes. In the analysis of covariance (ANCOVA), this is known as heterogeneity of regression. It is usual in ANCOVA to check whether regressions are

homogeneous. Similarly, in the hierarchical linear model it is advisable to check for each level-one variable in the fixed part whether it has a random slope.

Depending on the amount of data and the complexity of the model, estimating random slopes for each explanatory variable may be a time-consuming affair. A faster method for testing random slopes was proposed by Berkhof and Snijders (2001) and also by Verbeke and Molenberghs (2003). These tests do not require the complete ML or REML estimates for the random slope models. Suppose that some model, denoted by M_0 , has been estimated, and it is to be tested whether a random slope for some variable should be added to the model. Thus, M_0 is to be tested as the null hypothesis. The method proceeds as follows.

Recall from Section 4.7 that the estimation algorithms for the hierarchical linear model are iterative procedures. If the IGLS, RIGLS, or Fisher scoring algorithm is used, it is already possible to base a test of the random slope on the results of the first step of the algorithm, making it unnecessary to carry out the full estimation process with all iteration steps. To test whether some variable has a random slope, first estimate the parameters for model M_0 and then add this random slope to M_0 (it is sufficient to add only the variance parameter, without the corresponding covariance parameters). Now carry out just one step of the (R)IGLS or Fisher scoring algorithm, starting from the parameter estimates obtained when fitting M_0 . Denote the provisional slope variance estimate after this single step by $\tilde{\tau}^2$, with associated standard error (also after the single step) S.E.($\tilde{\tau}^2$). Then the t -ratio,

$$\frac{\tilde{\tau}^2}{\text{S.E.}(\tilde{\tau}^2)},$$

can be tested against the standard normal distribution. (Note that, as is explained in Berkhof and Snijders (2001), this test of a t -ratio for a variance parameter can be applied only to the result of the first step of the estimation algorithm and not to the final ML estimate. The main reason is that the standard error of the usual estimator, obtained on convergence of the algorithm, is to a certain extent proportional to the estimated variance component, in contrast to the standard error of the single-step estimate.)

Another question that is relevant when a random slope is specified for a variable X of which the between-group regression coefficient differs from the within-group coefficient, is whether the original variable X_{ij} or the within-group deviation variable $X_{ij} - \bar{X}_j$ should get the random slope. This was discussed in Section 5.3.1.

10.2.3 Explained variance

In the discussion in Section 7.1 of the concept of explained variance for multilevel models, it was mentioned that the fraction of explained variance at level one, R_1^2 , can be used as a misspecification diagnostic. If this fraction of explained variance decreases when a fixed effect is added to the model, this may be a sign of misspecification. A small decrease, however, may be a result of chance fluctuations. For reasonably large data sets, a decrease by a magnitude of 0.05 or more should be taken as a warning of possible misspecification.

10.3 Specification of the fixed part

Whether the fixed part contains the right variables is an equivocal question in almost all research in the social and behavioral sciences. For many dependent variables there is no

single correct explanation, but different points of view may give complementary and valid representations. The performance of a pupil at the final moment of compulsory education can be studied as a function of the characteristics of the child's family; as a function of the child's achievements in primary school; as a function of the child's personality; or as a function of the mental processes during the final examination. All these points of view are valid, taken alone and also combined. Therefore, what is the right set of variables in the fixed part depends in the first place on the domain in which the explanation is being considered.

In the second place, once the data collection is over one can only use variables based on the available data. Both of these considerations imply that some variable, say X_q , may be omitted from the fixed part of the model. Part or all of this variable will then be represented by the residuals in the random part. If the fixed part includes some variables that are correlated with X_q , then the effects of the variables included will take up some of the effect of the variable omitted. This is well known in regression analysis. It implies that (unless the variables are experimentally manipulated) one should be careful with interpretation and very reluctant to make causal interpretations. For the hierarchical linear model the consequences of omitting variables from the fixed part of the model are discussed in Raudenbush and Bryk (2002, Chapter 9).

In a multilevel design, one should be aware in any case of the possibility that supposed level-one effects are in reality, completely or partially, higher-level effects of aggregated variables. This was discussed in Section 10.2.1.

Transformation of explanatory or dependent variables

One important option for improving the specification of the fixed part of the hierarchical linear model is the transformation of explanatory variables. Aggregation to group means or to group standard deviations is one kind of transformation. Calculating products or other interaction variables is another kind. The importance given in multilevel modeling to cross-level interactions should not diminish the attention paid to the possibility of within-level interactions.

A third kind of transformation is the nonlinear transformation of variables. Such transformations may be applied to the dependent as well as the explanatory variables. Our treatment focuses on the latter, but the former possibility should not be forgotten. As examples of nonlinear effects of an explanatory variable X on Y , one may think of an effect which is always positive but levels off toward low or high values of X ; or a U-shaped effect, expressed by a function that first decreases and then increases again.

The simplest way to investigate the possible nonlinearity of the effect of a variable X is to conjecture some nonlinear transformation of X , such as X^2 , $\ln(X)$ (for positive variables), or $1/X$, add this effect to the fixed part of the model – retaining the original linear effect of X – and test the significance of the effect of this nonlinear transformed variable. Instead of adding just one nonlinear transformed variable, flexibility is increased by adding several transformations of the same variable X , because the linear combinations of these transformations will be able to represent quite a variety of shapes of the effect of X on Y .

Rather than conjecturing a specific good transformation, one may also choose the best one from within a family of transformations. A well-known family of transformations is the Box–Cox transformation (Box and Cox, 1964) defined by

$$x^\lambda = \begin{cases} \frac{x^\lambda - 1}{\lambda} & (\lambda \neq 0) \\ \ln(x) & (\lambda = 0), \end{cases} \quad (10.2)$$

where \ln denotes the natural logarithm and the parameter λ must be determined so as to give the best fit. An interesting feature of this family is that it shows how the logarithm is embedded in a family of power transformations. It is a family of transformations often applied to the dependent variable, in cases where there is skewness in the distribution of the residuals. This transformation is treated in Cook and Weisberg (1982, Chapter 5) and Atkinson (1985), and for transforming dependent variables in multilevel models by Hedges (1998, p. 506) and Goldstein et al. (2009).

Another convenient family of transformations for explanatory variables is constituted by spline functions. Spline functions are discussed in Section 15.2.2. More extensive discussions are given in Seber and Wild (1989, Section 9.5) and Fox (2008, Chapter 17); and for multilevel models in Goldstein (2011, Chapter 15) and Snijders and Berkhof (2008, Sections 3.8–3.11). We find quadratic splines simple to work with and quite effective. An example is given in Figure 8.1. A quadratic spline transformation of some variable X can be chosen as follows. Choose a value x_0 in the middle of the range of X (e.g., the grand mean of X) as a reference point for the square of X and choose a few other values x_1, \dots, x_K within the range of X . Usually, K is a low number, such as 1, 2, or 3. Higher values of K yield more flexibility to approximate many shapes of functions of X . Calculate, for each $k = 1, \dots, K$, the ‘half-squares’ $f_k(X)$, defined as 0 left of x_k and $(X - x_k)^2$ right of x_k :

$$f_k(X) = \begin{cases} 0 & (X \leq x_k) \\ (X - x_k)^2 & (X > x_k). \end{cases}$$

In the fixed part use the linear effect X , the quadratic effect $(X - x_0)^2$, and the effects of the ‘half squares’ $f_1(X), \dots, f_K(X)$. Together these functions can represent a wide variety of smooth functions of X , as is evident from Figure 8.1. If some of the $f_k(X)$ have non-significant effects they can be left out of the model, and by trial and error the choice of the so-called nodes x_k may be improved. Such an explorative procedure was used to obtain the functions displayed in this figure.

10.4 Specification of the random part

The specification of the random part has been discussed above and in Section 6.4. If certain variables are mistakenly omitted from the random part, the tests of their fixed coefficients may be unreliable. Therefore it is advisable to check the randomness of slopes of all variables of main interest, and not only those for which a random slope is theoretically expected.

The random part specification is directly linked to the structure of the covariance matrix of the observations. In Section 5.1.1 we saw that a random slope implies a heteroscedastic specification of the variances of the observations and of the covariance between level-one units in the same group (level-two unit). When different specifications of the random part

yield a similar structure for the covariance matrix, it will be empirically difficult or impossible to distinguish between them. But also a misspecification of the fixed part of the model can lead to a misspecification of the random part; sometimes an incorrect fixed part shows up in an unnecessarily complicated random part. For example, if an explanatory variable X with a reasonably high intraclass correlation has in reality a curvilinear (e.g., quadratic) effect without a random component, whereas it is specified as having a linear fixed and random effect, the excluded curvilinear fixed effect may show up in the shape of a significant random effect. The latter effect will then disappear when the correctly specified curvilinear effect is added to the fixed part of the model. This was observed in Example 8.2. The random slope variance of IQ was attenuated in this example when a curvilinear effect of IQ was considered.

10.4.1 Testing for heteroscedasticity

In the hierarchical linear model with random slopes, the *observations* are heteroscedastic because their variances depend on the explanatory variables, as expressed by equation (5.5). However, the *residuals* R_{ij} and U_{hj} are assumed to be homoscedastic, that is, to have constant variances.

In Chapter 8 it was explained that the hierarchical linear model can also represent models in which the level-one residuals have variances depending linearly or quadratically on an explanatory variable, say, X . Such a model can be specified by the technical device of giving this variable X a random slope at level one. Similarly, giving a level-two variable Z a random slope at level two leads to models for which the level-two random intercept variance depends on Z . Also random slope variances can be made to depend on some variable Z . Neglecting such types of heteroscedasticity may lead to incorrect hypotheses tests for variables which are associated with the variables responsible for this heteroscedasticity (X and Z in this paragraph). Checking for this type of heteroscedasticity is straightforward using the methods of Chapter 8. However, this requires variables to be available that are thought to be possibly associated with residual variances.

A different method, described in Raudenbush and Bryk (2002, Chapter 9), can be used to detect heteroscedasticity in the form of between-group differences in the level-one residual variance, without a specific connection to some explanatory variable. It is based on the estimated least squares residuals within each group, called the ordinary least squares (OLS) residuals. This method is applicable only if many (or all) groups are considerably larger than the number of level-one explanatory variables. The level-two explanatory variables are, for the moment, neglected. What follows applies to the groups for which $n_j - r - 1$ is not too small (say, 10 or more), where r is the number of level-one explanatory variables. For each of these groups separately an OLS regression is carried out with the level-one variables as explanatory variables. Denote by s_j^2 the resulting estimated residual variance for group j and by $df_j = n_j - r - 1$ the corresponding number of degrees of freedom. The weighted average of the logarithms,

$$ls_{\text{tot}} = \frac{\sum_j df_j \ln(s_j^2)}{\sum_j df_j}, \quad (10.3)$$

must be calculated. If the hierarchical linear model is well specified this weighted average will be close to the logarithm of the maximum likelihood estimate of σ^2 .

From the group-dependent residual variance s_j^2 a standardized residual dispersion measure can be calculated using the formula

$$d_j = \sqrt{\frac{df_j}{2}} \left\{ \ln(s_j^2) - ls_{\text{tot}} \right\}. \quad (10.4)$$

If the level-one model is well specified and the population level-one residual variance is the same in all groups, then the distribution of the values d_j is close to the standard normal distribution. The sum of squares,

$$H = \sum_j d_j^2, \quad (10.5)$$

can be used to test the constancy of the level-one residual variances. Its null distribution is chi-squared with $N' - 1$ degrees of freedom, where N' is the number of groups included in the summation.

If the within-groups degrees of freedom df_j are less than 10 for many or all groups, the null distribution of H is not chi-squared. Since the null distribution depends only on the values of df_j and not on any of the unknown parameters, it is feasible to obtain this null distribution by straightforward computer simulation. This can be carried out as follows: generate independent random variables V_j according to chi-squared distributions with df_j degrees of freedom, calculate $s_j^2 = V_j/df_j$, and apply equations (10.3), (10.4) and (10.5). The resulting value H is one random draw from the correct null distribution. Repeating this, say, 1,000 times gives a random sample from the null distribution with which one can compare the observed value from the real data set.

If this test yields a significant result, one can inspect the individual d_j values to investigate the pattern of heteroscedasticity. For example, it is possible that the heteroscedasticity is due to a few unusual level-two units for which d_j has a large absolute value. For other approaches for dealing with heteroscedasticity, see Section 10.4.2.

Example 10.1 Level-one heteroscedasticity.

The example of students' language performance used in Chapters 4, 5, and 8 is considered again. We investigate whether there is evidence of level-one heteroscedasticity where the explanatory variables at level one are IQ, SES, and gender, specifying the model as Model 1 in Table 8.1. Note that the nature of this heteroscedasticity test is such that the level-two variables included in the model do not matter. Those groups were used for which the residual degrees of freedom are at least 10. There were 133 such groups. The sum of squared standardized residual dispersions defined in (10.5) is $H = 155.5$, a chi-squared value with $df = 132$, yielding $p = 0.08$. Hence this test does not give evidence of heteroscedasticity.

Although nonsignificant, the result is close enough to significance that it is still worthwhile to try and look into this small deviation from homoscedasticity. The values d_j can be regarded as standard normal deviates in the case of homoscedasticity. The two schools with the largest absolute values had $d_j = -3.2$ and -3.8 , expressing low within-school variability; while the other values were all less than 2.5 in absolute value, which is of no concern. The two homogeneous schools were also those with the highest averages for the dependent variable, the language test. They were not particularly large or small, and had average compositions with respect to socio-economic status and IQ. Thus, they were outliers in two associated ways: high averages and low internal residual variability. The fact that the homoscedasticity test gave $p = 0.08$, not significant at the conventional level of 0.05, suggests, however, that these outliers are not serious.

An advantage of this test is that it is based only on the specification of the within-groups regression model. The level-two variables and the level-two random effects play no role at all, so what is checked here is purely the level-one specification. However, the null distributions of the d_j and of H mentioned above do depend on the normality of the level-one residuals. A heavier-tailed distribution for these residuals in itself will also lead to higher values of H , even if the residuals do have constant variance. Therefore, if H leads to a significant result, one should investigate the possible pattern of heteroscedasticity by inspecting the values of d_j , but one should also inspect the distribution of the OLS within-group residuals for normality.

10.4.2 What to do in case of heteroscedasticity

If there is evidence for heteroscedasticity, it may be possible to find variables accounting for the different values of the level-one residual variance. These could be level-one as well as level-two variables. Sometimes such variables can be identified on the basis of theoretical considerations. In addition, plots of d_j versus relevant level-two variables or plots of squared unstandardized residuals (see Section 10.5) may be informative for suggesting such variables. When there is a conjecture that the nonconstant residual variance is associated with a certain variable, one can apply the methods of Chapter 8 to test whether this is indeed the case, and fit a heteroscedastic model.

In some cases, a better approach to deal with heteroscedasticity is to apply a nonlinear transformation to the dependent variable, such as a square root or logarithmic transformation, as discussed in Section 10.3. How to choose transformations of the dependent variable in single-level models is discussed in Atkinson (1985), Cook and Weisberg (1982), and Goldstein et al. (2009). Transforming the dependent variable can be useful, for example, when the distribution of the dependent variable (or rather, of the residuals) is highly skewed.

When there is heteroscedasticity and the dependent variable has a small number of categories, another option is to use the multilevel ordered logit model for multiple ordered categories (Section 17.4) or to dichotomize the variable and apply multilevel logistic regression (Section 17.2).

10.5 Inspection of level-one residuals

A plethora of methods have been developed for the inspection of residuals in ordinary least squares regression; see, for example, Atkinson (1985) and Cook and Weisberg (1982, 1999). Inspection of residuals can be used, for example, to find outlying cases that have an undue high influence on the results of the statistical analysis, to check the specification of the fixed part of the model, to suggest transformations of the dependent or explanatory variables, or to point to heteroscedasticity.

These methods may likewise be applied to the hierarchical linear model, but some changes are necessary because of the more complex nature of the hierarchical linear model and the fact that there are several types of residuals. For example, in a random intercept model there is a level-one and also a level-two residual. Various methods of residual inspection for multilevel models were proposed by Hilden-Minton (1995). He noted that a problem in residual analysis for multilevel models is that the observations depend on

the level-one and level-two residual jointly, whereas for model checking it is desirable to consider these residuals separately. It turns out that level-one residuals can be estimated so that they are unconfounded by the level-two residuals, but the other way around is impossible.

Analysis of level-one residuals can be based on the OLS regressions calculated within each group separately. This was proposed by Hilden-Minton (1995). We made use of the within-group OLS regression in Section 10.4.1, where we remarked that an advantage of this approach is its being based only on the specification of the level-one model. Thus, it is possible to inspect estimated level-one residuals without confounding with level-two residuals or with level-two misspecification.

The estimated residuals are obtained by carrying out OLS regressions in each group separately, using only the level-one variables. It is advisable to omit the residuals for groups with low within-group residual degrees of freedom, because those residuals are quite unstable.

The variance of the OLS residuals depends on the explanatory variables. The OLS residuals can be standardized by dividing them by their standard deviation. This yields the *standardized OLS residuals*; see Atkinson (1985) or other texts on regression diagnostics. This provides us with two sets of level-one residuals: the raw OLS residuals and the standardized OLS residuals. The advantage of the standardized residuals is their constant variance; the disadvantage is that the standardization may distort the (possibly nonlinear) relations with explanatory variables.

Various possibilities for employing residuals to inspect model fit were discussed by Hilden-Minton (1995). For basic model checking we propose to use the level-one within-group OLS residuals in the following ways:

1. Plot the *unstandardized OLS residuals* against level-one explanatory variables to check for possible nonlinear fixed effects.

If there are in total a few hundred or more level-one units a raw plot may seem to exhibit just random scatter, in which case it is advisable to make a smoothed plot. The residual plot can be smoothed as follows.

Often the explanatory variable has a limited number of categories; not only when it is composed of a few ordered categories, but also, for example, when it is based on an underlying scale with integer values, possibly transformed, as in Example 10.2 below. Then the mean residual can be calculated for each category ('binning') and plotted together with vertical bars extending to twice the standard error of the mean within this category. (This standard error can be calculated in the usual way, without taking account of the nonconstant variance and the mutual dependence of the residuals.)

For continuous explanatory variables many statistical packages offer some method of smoothing, for example, a locally weighted scatterplot smoother abbreviated as *loess* or *lowess* (Cleveland, 1979). If such a smoother is not available, a simple and usually effective way of smoothing is the following construction of a simple moving average. Denote by x_1, \dots, x_M the ordered values of the explanatory variable under consideration and by r_1, \dots, r_M the corresponding values of the residuals. This means that x_i and r_i are values of the explanatory variable and the within-group residual for the same level-one unit, reordered so that $x_1 \leq x_2 \leq \dots \leq x_M$. (Units with identical x -values are ordered arbitrarily.) Since groups with a small number of units may have been omitted,

the number M may be less than the original total sample size. The smoothed residuals now are defined as averages of $2K$ consecutive values,

$$\bar{r}_i = \frac{1}{2K} \sum_{h=-K+1}^K r_{i+h},$$

for $i = K, \dots, M - K$. The value of K will depend on data and sample size; for example, if the total number of residuals M is at least 1,500, one could take moving averages of $2K = 100$ values.

One may add to the plot horizontal lines plotted at r equal to plus or minus twice the standard error of the mean of $2K$ values, that is, $r = \pm 2\sqrt{\hat{\sigma}^2/(2K)}$, where $\hat{\sigma}^2$ is the variance of the OLS residuals. This indicates roughly that values \bar{r}_i outside this band, that is, $|\bar{r}_i| > 2\sqrt{\hat{\sigma}^2/(2K)}$, may be considered to be relatively large.

2. Make a normal probability plot of the *standardized OLS residuals* to check the assumption of a normal distribution. This is done by plotting the values (z_i, \tilde{r}_i) where \tilde{r}_i is the standardized OLS residual and z_i the corresponding normal score (i.e., the expected value from the standard normal distribution according to the rank of \tilde{r}_i). Especially when this shows that the residual distribution has longer tails than the normal distribution, there is a danger of parameter estimates being unduly influenced by outlying level-one units.

The *squared standardized residuals* can also be plotted against explanatory variables to assess the assumption of level-one homoscedasticity. Here also, averaging within categories, or smoothing, can be very helpful in showing the patterns that may exist.

When a data exploration is carried out along these lines, this may suggest model improvements which can then be tested. One should realize that these tests are suggested by the data; if the tests use the same data that were used to suggest them, which is usual, this will lead to capitalization on chance and inflated probabilities of type I errors. The resulting improvements are convincing only if they are ‘very significant’ (e.g., $p < 0.01$ or $p < 0.001$). If the data set is large enough, it is preferable to employ cross-validation (see p. 126).

Example 10.2 Level-one residual inspection.

We continue Example 10.1, in which the data set also used in Chapters 4 and 5 is considered again, the explanatory variables at level one being IQ, SES, and gender, with an interaction between IQ and SES. Within-group OLS residuals were calculated for all groups with at least 10 within-group residual degrees of freedom.

The variables IQ and SES are both based on integer scales from which the mean was subtracted, so they have a limited number of categories. For each category, the mean residual was calculated and the standard error of this mean was calculated in the usual way. The mean residuals are plotted in Figure 10.1 for IQ categories containing 12 or more students; for SES categories were formed as pairs of adjacent values, because otherwise they would be too small.

The vertical lines indicate the intervals bounded by the mean residual plus or minus twice the standard error of the mean. For IQ the mean residuals exhibit a clear pattern, while for SES they do not. The left-hand figure suggests a nonlinear function of IQ having a local minimum for IQ between -2 and 0 and a local maximum for IQ somewhere about 2. There are few students with IQ values less

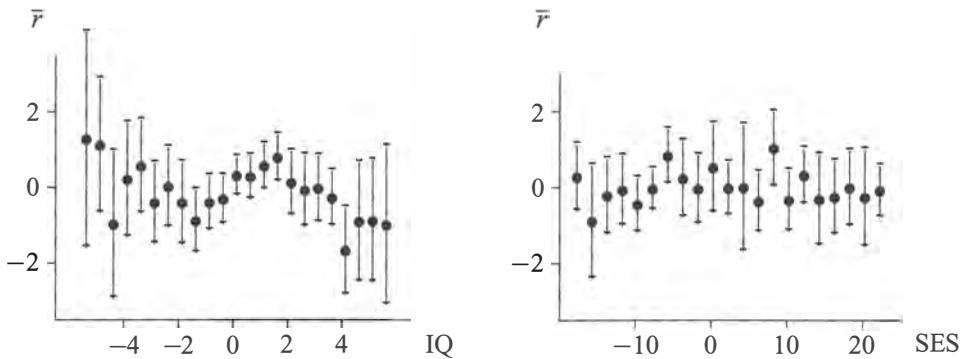


Figure 10.1: Mean level-one OLS residuals (with bars extending to twice the standard error of the mean) as function of IQ (left) and SES (right).

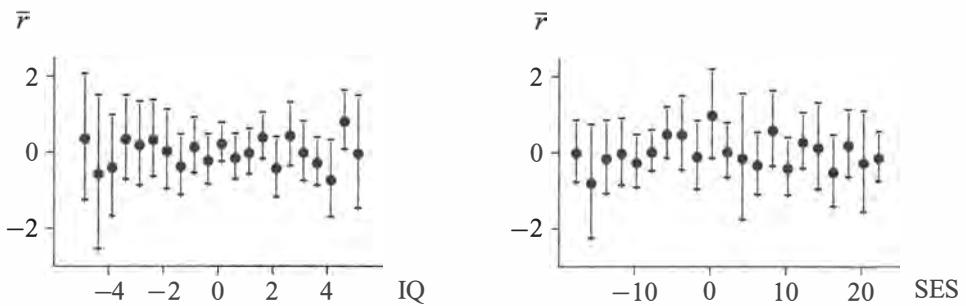


Figure 10.2: Mean level-one OLS residuals (with bars extending to twice the standard error of the mean) as function of IQ (left) and SES (right), for model with nonlinear effect of IQ.

than -4 or greater than $+4$, so in this IQ range the error bars are very wide and not very informative. Thus the figures point toward a nonlinear effect for IQ, in which the deviation from linearity has a local minimum for a negative IQ value and a local maximum for a positive IQ value. Examples of such functions are third-degree polynomials and quadratic spline functions with two nodes (cf. Section 15.2.2). The first option was explored by adding IQ^2 and IQ^3 to the fixed part. The second option was explored by adding IQ_-^2 and IQ_+^2 , as defined in (8.2), to the fixed part. The second option gave a much better model improvement and therefore was selected. When IQ_-^2 and IQ_+^2 are added to Model 1 of Table 8.1, which yields the same fixed part as Model 4 of Table 8.2, the deviance goes down by 54.9 ($df = 2, p < 0.00001$). This is strongly significant, so this nonlinear effect of IQ is convincing even though it was not hypothesized beforehand but suggested by the data. The mean residuals for the resulting model are graphed as functions of IQ and SES in Figure 10.2. These plots do not exhibit a remaining nonlinear effect for IQ.

A normal probability plot of the standardized residuals for the model that includes the nonlinear effect of IQ is given in Figure 10.3. The distribution looks quite normal except for the very low values, where the residuals are somewhat more strongly negative (i.e., larger in absolute value) than expected. However, this deviation from normality is rather small.

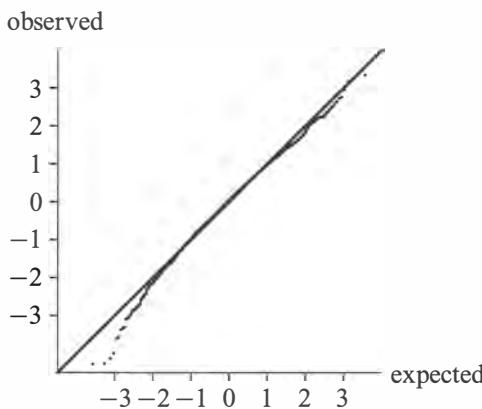


Figure 10.3: Normal probability plot of standardized level-one OLS residuals.

10.6 Residuals at level two

Estimated level-two residuals always are confounded with the estimated level-one residuals. Therefore one should check the specification of the level-one model before moving on to checking the specification at level two. The level-two specification can be checked by using the level-two residuals treated in this section and the influence measures treated in the next.

The empirical Bayes estimates (also called posterior means) of the level-two random effects, treated in Section 4.8, are the ‘estimates’ of the level-two random variables U_{hj} ($h = 0, 1, \dots, p$) in the model specification (10.1). These can be used as estimated level-two residuals. They can be standardized by dividing by their diagnostic standard error, as defined in Section 4.8.1.

Similarly to the level-one residuals, one may plot the unstandardized level-two residuals as a function of relevant level-two variables to check for possible nonlinearity; normal probability plots may be made of the standardized level-two residuals; and the squared standardized level-two residuals may be plotted as a function of level-two variables to check homoscedasticity. Smoothing the plots will be less often necessary because of the usually much smaller number of level-two units. If the plots contain outliers, one may inspect the corresponding level-two units to check whether anything unusual can be found.

Normal probability plots of standardized residuals are sensitive to both nonnormal residual distributions and misspecification of the fixed part of the model (Verbeke and Molenberghs, 2000, Section 7.8; Eberly and Thackeray, 2005). Therefore one should realize that what seems to be a deviation from normality might also be caused by an incorrect specification of the fixed effects.

Checking empirical Bayes residuals is discussed more extensively in Langford and Lewis (1998, Section 2.4). If the diagnostic variances differ considerably between level-two units, greater precision can be obtained by using the diagnostic variances also to weight the contributions of the standardized residuals to their cumulative distribution function; this was proposed by Lange and Ryan (1989).

Example 10.3 Level-two residual inspection.

The example above is continued now with the inspection of the level-two residuals. The model is specified with fixed effects of IQ and SES and their interaction; the school means of IQ and SES and their interaction; the two nonlinear functions of IQ; and gender. The random part at level two includes a random intercept and a random slope for IQ. (This is like Model 4 of Table 8.2, except that the level-one heteroscedasticity is left out.) Level-two empirical Bayes residuals are calculated for the intercept and for the slope of IQ; these may be called ‘posterior intercepts’ and ‘posterior slopes’, respectively. Figure 10.4 shows the unstandardized posterior intercepts as a function of the school averages of IQ and SES. Locally weighted scatterplot (‘lowess’) smoothers (Cleveland, 1979), as mentioned on p. 162, are added as nonlinear approximations to the averages of the residuals as a function of these school variables to provide a better view of how the average residuals depend on the school means. Figure 10.5 shows the unstandardized posterior slopes for IQ, as a function of the school averages of IQ and SES. The four plots show some weak patterns, but not enough to justify further model extensions.

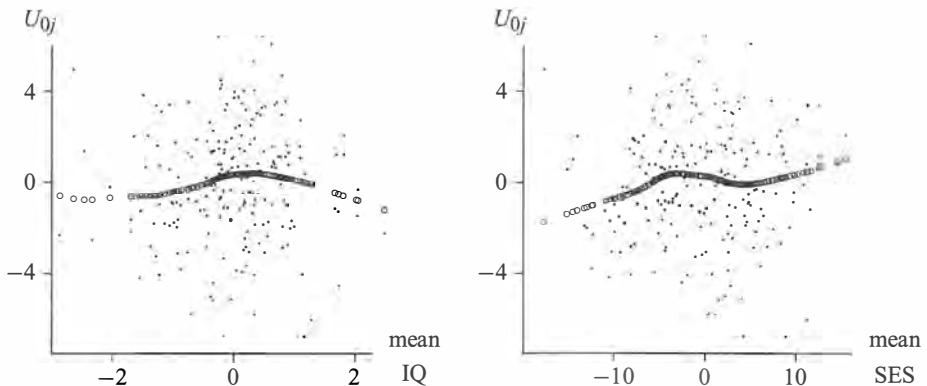


Figure 10.4: Posterior intercepts as function of (left) average IQ and (right) average SES for each school. Smooth lowess approximations are indicated by o.

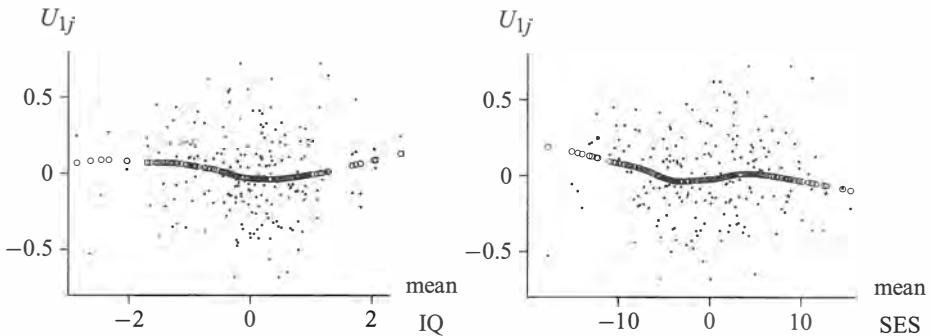
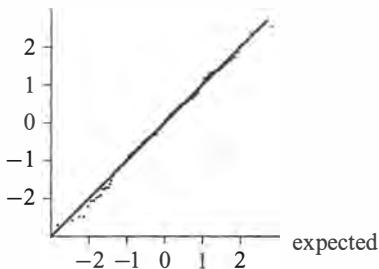


Figure 10.5: Posterior IQ slopes as function of (left) average IQ and (right) average SES for each school. Smooth lowess approximations are indicated by o.

In Figure 10.6 the normal probability plots of the standardized residuals are shown. These normal plots support the approximate normality of the level-two residuals.

observed



observed

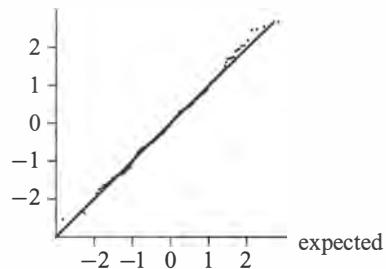


Figure 10.6: Normal probability plot of standardized level-two residuals:
(left) intercepts, (right) slopes.

10.7 Influence of level-two units

The diagnostic value of residuals can be supplemented by so-called *influence diagnostics*, which give an indication of the effect of certain parts of the data on the parameter estimates obtained. In this section we present diagnostics to investigate the influence of level-two units. These diagnostics show how strongly the parameter estimates are affected if unit j is deleted from the data set. The residuals treated here are based on the *deletion principle* (Atkinson, 1985), that is, they are obtained by comparing the data for a given unit with the values expected under the model, estimated from a data set from which this unit has been deleted. This treatment is based on Snijders and Berkhof (2008, Section 3), in which a modification of Lesaffre and Verbeke (1998) and of the corresponding section in the first edition of the current textbook was developed. Other influence measures for multilevel models may be found in Langford and Lewis (1998, Section 2.3).

For the definition of the diagnostics we shall need matrix notation, but readers unfamiliar with this notation can base their understanding on the verbal explanations.

Denote by $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_r)$ the vector of all parameters of the fixed part, consisting of the general intercept and all regression coefficients. Further, denote by $\hat{\gamma}$ the vector of estimates produced when using all the data and by $\hat{\Sigma}_F$ the covariance matrix of this vector of estimates (so the standard errors of the elements of γ are the square roots of the diagonal elements of $\hat{\Sigma}_F$). Denote the parameter estimate obtained if level-two unit j is deleted from the data by $\hat{\gamma}_{(-j)}$. Then unit j has a large influence if $\hat{\gamma}_{(-j)}$ differs much from $\hat{\gamma}$. This difference can be measured on the basis of the covariance matrix $\hat{\Sigma}_F$, because this matrix indicates the uncertainty that exists anyway in the elements of $\hat{\gamma}$. For increased diagnostic precision, a modified estimate $\hat{\Sigma}_{F(-j)}$ is used, based on the data from which group j has been deleted.

Unit j has a large impact on the parameter estimates if, for one or more of the individual regression coefficients $\hat{\gamma}_h$, the difference between $\hat{\gamma}_h$ and $\hat{\gamma}_{(-j)h}$ is not much smaller, or even larger, than the standard error of $\hat{\gamma}_h$. Given that the vector γ has a total of $r + 1$ elements, a standardized measure of the difference between the estimated fixed effects for the entire data set and the estimates for the data set excluding unit j is given by

$$C_j^{\text{OF}} = \frac{1}{r+1} (\hat{\gamma} - \hat{\gamma}_{(-j)})' \hat{\Sigma}_{F(-j)}^{-1} (\hat{\gamma} - \hat{\gamma}_{(-j)}).$$

This can be interpreted as the average squared deviation between the estimates with and those without unit j , where the deviations are measured proportional to the standard errors (and account is taken of the correlations between the parameter estimates). For example, if there is only one fixed parameter and the value of C_j^{OF} for some unit is 0.5, then leaving out this unit changes the parameter estimate by a deviation equal to $\sqrt{0.5} = 0.7$ times its standard error, which is quite appreciable.

This influence diagnostic is a direct analog of Cook's distance (e.g. Cook and Weisberg, 1982; Atkinson, 1985) for linear regression analysis. If the random part at level two is empty, then C_j^{OF} is equal to Cook's distance.

It may be quite time-consuming to calculate the estimates $\hat{\gamma}_{(-j)}$ for each group j . Since the focus here is on the diagnostic value of the influence statistic rather than on the precise estimation for the data set without group j , an approximation can be used instead. Pregibon (1981) proposed, in a different context, substituting the one-step estimator for $\hat{\gamma}_{(-j)}$. This is the estimator which starts from the estimate obtained for the full data set and then carries out a single step of the IGLS, RIGLS, or Fisher scoring algorithm. Denoting this one-step estimator by $\tilde{\gamma}_{(-j)}$ and the corresponding one-step estimator for the covariance matrix by $\tilde{\Sigma}_{F(-j)}$, we obtain the influence diagnostic

$$C_j^{\text{F}} = \frac{1}{r+1} (\hat{\gamma} - \tilde{\gamma}_{(-j)})' \tilde{\Sigma}_{F(-j)}^{-1} (\hat{\gamma} - \tilde{\gamma}_{(-j)}). \quad (10.6)$$

A similar influence diagnostic can be defined for the influence of group j on the estimates of the random part. Denote by φ the vector of all parameters of the random part and by q the number of elements of this vector. If the covariance matrix of the random effects is unrestricted, then φ contains the variances and covariances of the random slopes as well as the level-one residual variance. Given that there are p random slopes, this is a total of $q = (p+1)(p+2)/2 + 1$ parameters. Now define $\hat{\varphi}$ as the estimate based on the complete data set and $\hat{\Sigma}_R$ as the covariance matrix of this estimate (having on its diagonal the squared standard errors of the elements of $\hat{\varphi}$); and $\tilde{\varphi}_{(-j)}$ and $\tilde{\Sigma}_{R(-j)}$ as the one-step estimates when deleting level-two unit j . Then the diagnostic for the influence of unit j on the parameters of the random part is defined by

$$C_j^{\text{R}} = \frac{1}{q} (\hat{\varphi} - \tilde{\varphi}_{(-j)})' \tilde{\Sigma}_{R(-j)}^{-1} (\hat{\varphi} - \tilde{\varphi}_{(-j)}). \quad (10.7)$$

One can also consider the combined influence of group j on the parameters of the fixed and those of the random part of the model. Since the estimates $\hat{\gamma}$ and $\hat{\varphi}$ are approximately uncorrelated (Longford, 1987), such a combined influence diagnostic can be defined simply as the weighted average of the two previously defined diagnostics,

$$C_j = \frac{1}{r+q+1} ((r+1) C_j^{\text{F}} + q C_j^{\text{R}}). \quad (10.8)$$

Lesaffre and Verbeke (1998) propose local influence diagnostics which are closely related to those proposed here.¹

¹ Lesaffre and Verbeke's diagnostics are very close to diagnostics (10.6), (10.7), and (10.8), multiplied by twice the number of parameters, but using different approximations. We have opted for the diagnostics proposed here because their values are on the same scale as the well-known Cook's distance for the linear regression model, and because the one-step estimates can be conveniently calculated when one uses software based on the Fisher scoring or (R)IGLS algorithms.

Whether a group has a large influence on the parameter estimates for the whole data set depends on two things. The first is the *leverage* of the group, that is, the potential of this group to influence the parameter estimates because of its size n_j and the values of the explanatory variables in this group. Groups with a large size and with strongly dispersed values of the explanatory variables have a high leverage. The second is the extent to which this group fits the model as defined by (or estimated from) the other groups, which is closely related to the values of the residuals in this group. A poorly fitting group that has low leverage, for example, because of its small size, will not strongly affect the parameter estimates. Likewise, a group with high leverage will not strongly affect the parameter estimates if it has deletion residuals very close to 0.

If the model fits well and the explanatory variables are approximately randomly distributed across the groups, then the expected value of diagnostics (10.6), (10.7), and (10.8) is roughly proportional to the group size n_j . A plot of these diagnostics as a function of n_j will reveal whether some of the groups influence the fixed parameter estimates more strongly than should be expected on the basis of group size.

The fit of the level-two units to the model can be measured by the *deletion standardized multivariate residual* for each unit. This measure was also proposed by Snijders and Berkhof (2008) as a variant of the ‘squared length of the residual’ of Lesaffre and Verbeke (1998), modified by following the principle of deletion residuals. It is defined as follows. The predicted value for observation Y_{ij} on the basis of the fixed part, minimizing the influence of the data in unit j on this prediction, is given by

$$\tilde{Y}_{ij} = \tilde{\gamma}_{0(-j)} + \sum_{h=1}^r \tilde{\gamma}_{h(-j)} x_{hij}.$$

The deletion multivariate residual for group j is the vector of deviations between the observations and these predicted values,

$$D_j = \begin{pmatrix} Y_{1j} - \tilde{Y}_{1j} \\ Y_{2j} - \tilde{Y}_{2j} \\ \vdots \\ Y_{n_j j} - \tilde{Y}_{n_j j} \end{pmatrix}. \quad (10.9)$$

This residual can be standardized on the basis of the covariance matrix of the vector of all observations in group j . For the hierarchical linear model with one random slope, for example, the elements of this covariance matrix are given by (5.5) and (5.6). Denote, for a general model specification, the covariance matrix of the observations in group j by

$$\Sigma(Y_j) = \text{cov}(Y_j).$$

This covariance matrix is a function of the parameters of the random part of the model. Substituting estimated parameters, again obtained by the deletion principle to minimize the influence of the outcomes observed for group j itself, yields the estimated covariance matrix, $\tilde{\Sigma}(Y_j)$. Now the standardized multivariate residual is defined by

$$S_j^2 = D_j' \left(\tilde{\Sigma}(Y_j) \right)^{-1} D_j. \quad (10.10)$$

This approach can also be used to define standardized deletion level-one residuals as

$$\tilde{r}_{ij} = \frac{Y_{ij} - \hat{Y}_{ij}}{\widetilde{\text{var}}(Y_{ij})}, \quad (10.11)$$

where $\widetilde{\text{var}}(Y_{ij})$ is the i th diagonal element of the covariance matrix $\tilde{\Sigma}(Y_j)$.

Unlike the level-one OLS residuals of Section 10.5, these definitions can also be applied to models with heteroscedasticity at level one (see Chapter 8).

The standardized residual can be interpreted as the sum of squares of the multivariate residual for group j after transforming this multivariate residual to a vector of uncorrelated elements each with unit variance. If the model is correctly specified and the number of groups is large enough for the parameter estimates to be quite precise, then S_j^2 has a chi-squared distribution with n_j degrees of freedom. This distribution can be used to test the value of S_j^2 and thus investigate the fit of this group to the hierarchical linear model as defined by the total data set. (This proposal was also made by Waternaux et al., 1989.) Some caution should be exercised with this test, however, because among all the groups some ‘significant’ results can be expected to occur by chance even for a well-specified model. For example, if there are 200 groups and the model is correctly specified, it may be expected that S_j^2 has a significantly large value at the 0.05 significance level for ten groups, and at the 0.01 level for two groups. One can apply here the Bonferroni correction for multiple testing, which implies that the model specification is doubtful if the smallest p -value of the standardized multivariate residuals is less than $0.05/N$, where N is the total number of groups.

Values of C_j are assessed only in comparison with the other groups, while values of S_j^2 can be assessed on the basis of the chi-squared distribution. When checking the model specification, one should worry mainly about groups j for which the influence diagnostic C_j is large compared to the other groups, and simultaneously there is a poor fit as measured by the standardized multivariate residual S_j^2 . For such groups one should investigate whether there is anything particular. For example, there may be data errors or the group may not really belong to the population investigated. This investigation may, of course, also lead to a model improvement.

Example 10.4 Influence of level-two units.

We continue checking the model for the language test score which was also investigated in Example 10.3. Recall that this is Model 4 of Table 8.2, but without the level-one heteroscedasticity. The 20 largest influence diagnostics (10.8) are presented in Table 10.1 together with the p -values of the standardized multivariate residuals (10.10). When assessing the p -values one should realize that, since these are the most influential schools, this set will contain some schools that differ relatively strongly from the others and therefore have relatively low p -values.

The school with the largest influence has a good fit ($p = 0.293$). School 15 fits very poorly, while schools 18, 52, 107, 122, and 187 have a moderately poor fit. The Bonferroni-corrected bound for the smallest p -value is $0.05/211 = 0.00024$. The poorest fitting school has a p -value of 0.00018, smaller than 0.00024, which suggests indeed that the fit is unsatisfactory, although not to a very strong extent.

As a further model improvement, the heteroscedasticity as a function of IQ was considered, leading to exactly Model 4 of Table 8.2. The 20 largest influence diagnostics for this model are presented in Table 10.2.

The influence diagnostic of school 213 has considerably increased; but this school does not have a very small number of students in the data set ($n_j = 19$) and it does not have a very poor fit to the

Table 10.1: The 20 largest influence diagnostics.

School	n_j	C_j	p_j
182	9	0.053	0.293
107	17	0.032	0.014
229	9	0.028	0.115
14	21	0.027	0.272
218	24	0.026	0.774
52	21	0.025	0.024
213	19	0.025	0.194
170	27	0.021	0.194
67	26	0.017	0.139
18	24	0.016	0.003
117	27	0.014	0.987
153	22	0.013	0.845
187	26	0.013	0.022
230	21	0.012	0.363
15	8	0.012	0.00018
256	10	0.012	0.299
122	23	0.012	0.005
50	24	0.011	0.313
101	23	0.011	0.082
214	21	0.011	0.546

rest of the data: $p = 0.010$ is not alarming giving that it is observed for the school selected as the one with the strongest influence. The fit of school 15 has improved, with p going from 0.00018 to 0.00049, and now exceeds the Bonferroni-corrected critical value of 0.00024. Therefore, the results shown in Table 10.2 are quite reassuring.

There is one school with a smaller p -value for the standardized multivariate residual. This is school 108, with $n_j = 9$ students and $p_j = 0.00008$. This is less than 0.00024 and therefore does lead to significance, for the poorest fitting school, with the Bonferroni correction. Inspection shows that the poor fit is caused by the fact that two out of the nine pupils in this school have very low values on the language test (14 and 17) while not having unusual values on IQ or SES. Since this school does not have a large influence ($C_j = 0.008$), however, the fact that it does not fit well is of no real concern.

A look at the data for school 108 reveals that the predicted values \hat{Y}_{ij} for this school are quite homogeneous and close to the average, but two of the nine outcome values Y_{ij} are very low. Correspondingly, the standardized deletion level-one residuals (10.5) are quite small, -3.0 and -4.3 . When these two students are deleted from the data set, the parameter estimates hardly change, but the lowest p -value for the standardized multivariate residuals is no longer significant when the Bonferroni correction is applied. Analyzing this data set while excluding the two schools with the poorest fit, 15 and 108, yields practically the same estimates as the full data set.

To conclude, the investigation of the influence statistics led to an improved fit in line with the heteroscedasticity models of Chapter 8, and to the detection of two deviant cases in the data set. These two cases out of a total of 3,758, however, did not have any noticeable influence on the parameter estimates.

Table 10.2: The 20 largest influence diagnostics for the extended model.

School	n_j	C_j	p_j
213	19	0.094	0.010
182	9	0.049	0.352
107	17	0.041	0.006
187	26	0.035	0.009
52	21	0.028	0.028
218	24	0.025	0.523
14	21	0.024	0.147
229	9	0.016	0.175
67	26	0.016	0.141
122	23	0.016	0.004
18	24	0.015	0.003
230	21	0.015	0.391
169	30	0.014	0.390
170	27	0.013	0.289
144	16	0.013	0.046
117	27	0.013	0.988
40	25	0.012	0.040
153	22	0.012	0.788
15	8	0.011	0.00049
202	14	0.010	0.511

10.8 More general distributional assumptions

If it is suspected that the normality and homoscedasticity assumptions of the hierarchical linear model are not satisfied, one could employ methods based on less restrictive model assumptions. This is an area of active research.

One approach is to derive estimators that are robust in the sense that the influence of individual cases or higher-level units is bounded. Such estimators were proposed by Richardson (1997), Yau and Kuk (2002), Dueck and Lohr (2005), and Copt and Victoria-Feser (2006). See also Demidenko (2004, Section 4.4) and Wu (2010, Chapter 9).

Another approach is to postulate distributions for residuals with heavier tails than the normal distribution; this will reduce influence for outliers. Models with t -distributed residuals were developed by Seltzer (1993), Seltzer et al. (1996), Pinheiro et al. (2001), Seltzer and Choi (2003), and Staudenmayer et al. (2009); and with more general residual distributions by Chen et al. (2002) and Burr and Doss (2005). Verbeke and Lesaffre (1996) propose mixtures of normal distributions for the residuals; see also Verbeke and Molenberghs (2000). Alonso et al. (2010) discuss the fact that greater difficulties may arise in the interpretation of estimated random effect variances for such methods.

Kasim and Raudenbush (1998) develop methods for models with heterogeneous level-one variances that are not (as in Chapter 8) a function of observed level-two variables.

For these models as well as for those assuming heavy-tailed distributions for higher-level residuals, computer-intensive methods such as Gibbs sampling are very useful; see Gelman et al. (2004, Chapter 15) and, more specifically, Congdon (2010, Sections 5.4, 6.5 and 10.4).

A different but very practical approach is provided by the latent class models of Section 12.3. These models approximate arbitrary distributions of random effects by discrete distributions and therefore can be used in principle for arbitrary random effect distributions.

‘Sandwich’ estimators for standard errors, also called cluster-robust standard errors, make no assumptions about the distributional shape of the random coefficients. Verbeke and Lesaffre (1997) and Yuan and Bentler (2002) proposed sandwich estimators to provide standard errors applicable for nonnormally distributed random effects. They are treated in Section 12.2. However, sandwich estimators do require large enough numbers of units at the highest level; see Verbeke and Lesaffre (1997), Maas and Hox (2004), and the further discussion in Section 12.2.

10.9 Glommary

Specification of the hierarchical linear model. This is embodied in the choice of the dependent variable, the choice of levels, of variables with fixed effects, and of variables with random effects with possible heteroscedasticity.

Assumptions of the hierarchical linear model. These are the linearity of the expected value of the dependent variable as a function of the explanatory variables, expressed by (10.1); the residuals at all levels having expected values 0 (independently of the predictor variables), constant variances (unless a heteroscedastic model is employed), and a normal distribution; and independence of the residuals across levels.

Model misspecification. This is a deviation between the process or ‘mechanism’ that generated the data and the specification of the model used to analyze the data. Misspecification can be serious because interesting conclusions may be masked, and erroneous inferences are possible – the ‘interesting’ and ‘nuisance’ aspects discussed in Chapter 2.

The logic of the hierarchical linear model. Following this can be useful in specifying the model: specify the levels correctly; consider whether variables with fixed effects may have different within-group and between-group effects (a choice with even more possibilities if there are three or more levels); ask similar questions for interaction effects; consider whether variables should have random slopes, and whether there might be heteroscedasticity. A considerable decrease of the explained variance (Section 7.1) when adding a fixed effect to the model is also a sign of potential misspecification of the model.

Transformations. Transforming the dependent as well as the explanatory variables can be helpful, for example, when variables have skewed distributions.

Box–Cox transformation. This is a family of transformations including powers (such as squares and square roots) and the logarithm. All these transformations are embedded

in a one-parameter family so that the choice of a good power or logarithmic transformation is implemented by choosing this parameter. This is a useful family of transformations for dependent as well as explanatory variables.

Spline functions. Polynomial functions (e.g., quadratic or cubic polynomials) between certain points (the ‘knots’ of the spline) linked smoothly at these knots. These are useful transformations for the explanatory variables.

Working upward. This is often a good way to check a multilevel model: first check the level-one model, then the level-two model (and so on, if there are higher levels). This is because the level-one model specification can be checked without confounding by the level-two model specification, but not the other way around.

Level-one residuals. The difference between predicted and observed values for the basic outcomes; to define these, a particular choice for ‘predicted’ must be made. A simple choice is predictions by OLS regressions for each level-two unit (‘group’) separately. This leads to the OLS within-group residuals. These can be used to test the specification of the within-group model: choice of fixed effects and homoscedasticity. To check the linearity of effects, smoothed plots of the level-one residuals can be used.

Normal probability plot. This is a plot of the residuals as a function of expected values from a normal distribution (also called a normal quantile–quantile or normal QQ plot). The plot will be approximately linear in the case of normal distributions. It can be employed to check the normality of level-one and level-two residuals.

Level-two residuals. These can be defined as empirical Bayes estimates, or posterior means, of the level-two random effects; since these have expected value 0, these estimates themselves are also the residuals. Empirical Bayes estimates were defined in Section 4.8. Here also plots can be made of the smoothed residuals as a function of relevant level-two explanatory variables to check linearity, as well as normal probability plots to check normality.

Leverage. The potential of a part of the data, for example, a level-two unit, to influence the parameter estimates because of its size and the values of the explanatory variables in this part of the data set.

Goodness of fit. This is how well the data, or a part of the data, corresponds to the model. The residuals are an important expression of this: residuals close to 0 point to a good fit.

Outliers. Data points that do not correspond well to the model, as will be indicated by large residuals.

Influence diagnostics. Statistics expressing influence of parts of the data set on the parameter estimates. In this chapter we considered influence diagnostics for two-level units. Higher-level units potentially have larger influence than level-one units, because a higher-level unit contains more information. Large influence is the combined result of large leverage and poor or mediocre fit. If it is found that some level-two units have a high influence then it will be important to assess to what extent

this is due to large leverage and poor fit: we mostly do not mind if a large unit with a reasonable to medium fit has a strong influence.

Deletion diagnostics. Diagnostics based on deleting part of the data, re-estimating the parameters, and assessing how strongly these parameter estimates differ from those obtained for the full data. This is a general principle for constructing residuals; in this chapter we followed an approximation to this principle (following it completely may be time-consuming).

Cook's distance. A well-known influence diagnostic for linear regression models. The level-two influence diagnostics treated in this chapter are similar in spirit to Cook's distance, but now either applied separately to the fixed coefficients and the parameters of the random part, or to all parameters of the hierarchical linear model together.

Multivariate residual. This is the vector of deviations between the observations and predicted values for a given level-two unit, and can be standardized by its covariance matrix, estimated according to the deletion principle.

Standardized deletion level-one residuals. These can be defined in a corresponding way, but now for the individual level-one units. In contrast to the OLS within-group residuals, these residuals can also be defined for models with heteroscedasticity at level one, and are meaningful even for small groups.

Bonferroni correction. This is a simple, generally applicable way of dealing with multiple testing. If several hypotheses are tested simultaneously at significance level α_0 , the probability that *at least one* of the null hypotheses is erroneously rejected will be larger than α_0 . If the number of tested hypotheses is K , the simultaneous probability of this last event cannot be larger than $K \times \alpha_0$. Therefore, the Bonferroni correction consists of testing each hypothesis at level α/K , so that the probability of erroneously rejecting at least one null hypothesis is no more than α . This is justified if the K tests should indeed be regarded as one simultaneous test; otherwise it will lead to an unnecessarily conservative test. In this chapter, the Bonferroni correction was applied to testing whether a given level-two unit fits the model; this should be done for all level-two units jointly, which makes this correction appropriate.

Robust estimators. Estimators that retain good performance for data that do not satisfy the default model assumptions very well. Different types of robust estimators have been developed, for example, incorporating robustness against outliers or against misspecification of the distribution of the random effects.

11

Designing Multilevel Studies

Up to now it has been assumed that the researcher wishes to test interesting theories on hierarchically structured systems (or phenomena that can be thought of as having a hierarchical structure, such as repeated data) on available data. Or that multilevel data exist, and that one wishes to explore the structure of the data. This, of course, is the other way around. Normally a theory (or a practical problem that has to be investigated) will direct the design of the study and the data to be collected. This chapter focuses on one aspect of this research design, namely, the sample sizes. Sample size questions in multilevel studies have also been treated by Snijders and Bosker (1993), Mok (1995), Cohen (1998), Donner and Klar (2000), Rotondi and Donner (2009), and others. Snijders (2001, 2005) and Cohen (2005) provide primers for this topic.

Another aspect, the allocation of treatments to subjects within groups, is discussed in Raudenbush (1997) and by Moerbeek and Teerenstra (2011). Raudenbush et al. (2007) compare matching and covariance adjustment in the case of group-randomized experiments. Hedges and Hedberg (2007) treat the case where sample size is a function of the intraclass correlation and covariate effects, while Rotondi and Donner (2009) provide an approach to sample size estimation in which the distribution of the intraclass correlation varies. Hedeker et al. (1999) as well as Raudenbush and Liu (2001) present methods for sample size determination for longitudinal data analyzed by multilevel methods.

This chapter shows how to choose sample sizes that will yield high power for testing, or (equivalently) small standard errors for estimating, certain parameters in two-level designs, given financial and practical constraints. A problem in the practical application of these methods is that sample sizes that are optimal, for example, for testing some cross-level interaction effect are not necessarily optimal, for example, for estimating the intraclass correlation. The fact that optimality depends on one's objectives, however, is a general problem of life that cannot be solved by this textbook. If one wishes to design a good multilevel study it is advisable to determine first the primary objective of the study, express this objective in a parameter to be tested or estimated, and then choose sample sizes for which this parameter can be estimated with a small standard error, given financial, statistical, and other practical constraints. Sometimes it is possible to check, in addition, whether also for some other parameters (corresponding to secondary objectives), these sample sizes yield acceptably low standard errors.

A relevant general remark is that the sample size at the highest level is usually the most restrictive element in the design. For example, a two-level design with 10 groups, that is, a macro-level sample size of 10, is at least as uncomfortable as a single-level design with a sample size of 10. Requirements on the sample size at the highest level, for a hierarchical linear model with q explanatory variables at this level, are at least as stringent as requirements on the sample size in a single-level design with q explanatory variables.

OVERVIEW OF THE CHAPTER

Before discussing the issue of power in multilevel designs in more detail, we make some general introductory remarks on the power of statistical tests and the factors that determine power in single-level models: sample size, effect size, and significance level. Then two not too complicated situations are considered: the design of a study to estimate a population mean whilst using a two-stage sampling design, and a design to determine the number of measurements per level-two unit to get a reliable estimate for those level-two units. Next, we introduce the problem of budget constraints when determining optimal sample sizes at both levels of the hierarchy, and present a series of examples of sample size calculations when one wishes to estimate associations between variables in a multilevel situation. The issue whether one should randomize groups or individuals within groups when conducting an experiment is then discussed. Finally, we treat sample size issues in cases where one is interested in precisely estimating the intraclass correlation coefficient or one of the variance components.

Level-two units will be referred to as ‘groups’ or ‘clusters’.

11.1 Some introductory notes on power

When a researcher is designing a multistage sampling scheme, for example, to assess the effects of schools on the achievement of students, or to test the hypothesis that citizens in impoverished neighborhoods are more often victims of crime than other citizens, important decisions must be made with respect to the sample sizes at the various levels. For the two-level design in the first example the question may be phrased as follows: should one investigate many schools with a few students per school or a few schools with many students per school? Or, for the second example: should we sample many neighborhoods with only a few citizens per neighborhood or many citizens per neighborhood and only a few neighborhoods? In both cases we assume, of course, that there are budgetary constraints on the research being conducted. To phrase this question more generally and more precisely: how should researchers choose sample sizes at the macro and micro level in order to ensure a desired level of power given a relevant (hypothesized) effect size and a chosen significance level? The average micro-level sample size per macro-level unit will be denoted by n and the macro-level sample size by N . In practice the sizes of the macro-level units will usually be variable (even if only because of unintentionally missing data), but for calculations of desired sample sizes it normally is adequate to use approximations based on the assumption of constant ‘group’ sizes.

A general introduction to power analysis can be found in the standard work by Cohen (1988), or, for a quick introduction, Cohen’s (1992) power primer. The basic idea is that we

would like to find support for a research hypothesis (H_1) stating that a certain effect exists, and therefore we test a null hypothesis about the absence of this effect (H_0) using a sample from the population of interest. The significance level α represents the risk of mistakenly rejecting H_0 . This mistake is known as a type I error. In addition, β denotes the risk of disappointingly not rejecting H_0 , when the effect does exist in the population. This mistake is known as a type II error. The statistical power of a significance test is the probability of rejecting H_0 given the effect size in the population, the significance level α , and the sample size and study design. Power is therefore given by $1 - \beta$. As a rule of thumb, Cohen suggests that power is moderate when it is 0.50 and high when it is at least 0.80. Power increases as α increases, and also as the sample size and/or the effect size increase. The effect size can be conceived as the researcher's idea about 'the degree to which the null hypothesis is believed to be false' (Cohen, 1992, p. 156).

We suppose that the effect size is expressed by some parameter γ that can be estimated with a certain standard error, denoted by $S.E.(\hat{\gamma})$. Bear in mind that the size of the standard error is a monotone decreasing function of the sample size: the larger the sample size the smaller the standard error! In most single-level designs, the standard error of estimation is inversely proportional (or roughly so) to the square root of sample size.

The relation between effect size, power, significance level, and sample size can be presented in one formula. This formula is an approximation that is valid for practical use when the test in question is a one-sided t -test for γ with a reasonably large number of degrees of freedom (say, $df \geq 10$). Recall that the test statistic for the t -test can be expressed by the ratio $t = \hat{\gamma}/S.E.(\hat{\gamma})$. The formula is

$$\frac{\text{effect size}}{\text{standard error}} \approx z_{1-\alpha} + z_{1-\beta} = z_{1-\alpha} - z_\beta, \quad (11.1)$$

where $z_{1-\alpha}$, $z_{1-\beta}$, and z_β are the z -scores (values from the standard normal distribution) associated with the cumulative probability values indicated. If, for instance, α is chosen at 0.05 and $1 - \beta$ at 0.80 (so that $\beta = 0.20$), and an effect size of 0.50 is what we expect, then we can derive that we are searching for a minimum sample size that satisfies

$$\frac{0.50}{\text{standard error}} \leq \frac{1.64 + 0.84}{1.64 + 0.84} = 0.20.$$

Formula (11.1) contains four 'unknowns'. This means that if three of these are given, then we can compute the fourth. In most applications that we have in mind, the significance level α is given and the effect size is hypothetically considered (or guessed) to have a given value; either the standard error is also known and the power $1 - \beta$ is calculated, or the intended power is known and the standard error calculated. Given this standard error, we can then try to calculate the required sample size.

For many types of design one can choose the sample size necessary to achieve a certain level of power on the basis of Cohen's work. For nested designs, however, there are two kinds of sample size: the sample size of the micro-units within each macro-unit n and the sample size of the macro-units N , with $N \times n$ being the total sample size for the micro-units.

Thus, more units are being observed at the micro level than at the macro level, therefore the tradeoff between significance level and power can also differ between the levels. When one is interested in assessing the effect of a macro-level variable on a micro-level outcome whilst controlling for a set of micro-level covariates, one could employ different

significance levels at the different levels. In this case one may, for example, set α for testing effects at the micro level to be as small as 0.01, but for effects of the macro-level variable at 0.05, in order to obtain sufficient power.

11.2 Estimating a population mean

The simplest case of a multilevel study occurs when one wishes to estimate a population mean for a certain variable of interest (e.g., income, age, literacy), and one is willing to use the fact that the respondents are regionally clustered. This makes sense, since if one is interviewing people it is a lot cheaper to sample, say, 100 regions and then interview 10 persons per region, than to randomly sample 1,000 persons that may live scattered all over a country. In educational assessment studies it is of course more cost-efficient to sample a number of schools and then to take a subsample of students within each school, than to sample students completely at random (ignoring their being clustered in schools).

Cochran (1977, Chapter 9) provides formulas to calculate desired sample sizes for such two-stage sampling. On p. 242, he defines the *design effect* for a two-stage sample, which is the factor by which the variance of an estimate (which is the square of the standard error of this estimate) is increased because of using a two-stage sample rather than a simple random sample with the same total sample size. Since estimation variance for a simple random sample is inversely proportional to total sample size, the design effect is also the factor by which total sample size needs to be increased to achieve the same estimation variance (or, equivalently, the same standard error) as a simple random sample of the given size. In Section 3.4 we saw that the design effect is given by

$$\text{design effect} = 1 + (n - 1) \rho_I, \quad (3.17)$$

where n is the average sample size in the second stage of the sample and ρ_I is the intra-class correlation. Now we can use (3.17) to calculate required sample sizes for two-stage samples.

Example 11.1 Assessing mathematics achievement.

Consider an international assessment study on mathematics achievement in secondary schools. The mathematics achievement variable has unit variance, and within each country the mean should be estimated with a standard error of 0.02. If a simple random sample of size n is considered, it can readily be deduced from the well-known formula

$$\text{standard error} = \frac{\text{standard deviation}}{\sqrt{n}}$$

that the sample size should be $n = 1/(0.02^2) = 2,500$.

What will happen to the standard error if a two-stage sampling scheme is employed (first schools then students), if the intraclass correlation is 0.20, and assuming that there are no direct extra budgetary consequences of sampling schools (this might be the case where one is estimating costs, when the standard errors are imposed by some international board)?

The design effect, when one opts for sampling 30 students per sampled school, is now $1 + (30 - 1) \times 0.20 = 6.8$. So instead of having a simple random sample size of 2,500 one should use a two-stage sample size of $2,500 \times 6.8 = 17,000$ in order to achieve the same precision.

Suppose further that one also wishes to test whether the average maths achievement in the country exceeds some predetermined value. The power in this design, given this sample size, depends on the effect size and α . The effect size here is the difference between the actual average achievement and the tested predetermined value. Let us assume that α is set at 0.01 and that one wishes to know the effect size for which the power is as large as $\beta = 0.80$. According to (11.1), this will be the case when the effect size is at least $0.02 \times (2.33 + 0.84) = 0.063$.

11.3 Measurement of subjects

In Section 3.5 the situation was discussed where the macro-unit may be, for example, an individual subject, a school, or a firm, and the mean $\mu + U_j$ is regarded as the ‘true score’ of unit j which is to be measured. The level-one deviation R_{ij} is then regarded as measurement error. This means that, whereas in the previous section the question was how to measure the overall population mean, now we are interested in measuring the mean of a level-two unit, which we shall refer to as a *subject*. The observations Y_{ij} can then be regarded as *parallel test items* for measuring subject j . Suppose we wish to use an unbiased estimate, so that the posterior mean of Section 4.8 is excluded (because it is biased toward the population mean; see p. 63). The true score for subject j therefore will be measured by the observed mean, \bar{Y}_j .

In equation (3.21) the reliability of an estimate defined as the mean of n_j measurements was given by

$$\lambda_j = \frac{n_j \rho_l}{1 + (n_j - 1) \rho_l}.$$

Now suppose that it is desired to have a reliability of at least a given value λ_0 . The equation for the reliability implies that this requires at least

$$n_{\min} = \frac{\lambda_0 (1 - \rho_l)}{(1 - \lambda_0) \rho_l} \quad (11.2)$$

measurements for each subject.

Example 11.2 Hormone measurement.

Consider individual persons as subjects, for whom the concentration of some hormone is to be measured from saliva samples. Suppose it is known that the intraclass correlation of these measurements is 0.40, so that the reliability of an individual measurement is 0.40 (cf. equation (3.19)). When an aggregate measurement based on several saliva samples is to be made with a reliability of $\lambda_0 = 0.80$, this requires at least $n_{\min} = (0.80 \times 0.60) / (0.20 \times 0.40) = 6$ measurements per subject.

11.4 Estimating association between variables

When one is interested in the effect of an explanatory variable on some dependent variable in a two-level hierarchical linear model, two situations are possible: both variables are micro-level variables, or the dependent variable is a micro-level variable whereas the

explanatory variable is defined at the macro level. Generally, in both cases, standard errors are inversely proportional to \sqrt{N} , the square root of the sampled number of macro-units, which implies that taking as many macro-units as possible (while n , the sample size within each sampled macro-unit, is constant) reduces the standard errors in the usual way, which in turn has a positive effect on the power of the statistical tests. Now let us suppose that the researcher is restricted to a given total sample size $M = N \times n$, but that he or she is free in choosing either N large with n small or N small with n large.

Snijders and Bosker (1993) developed formulas for standard errors of fixed effects in hierarchical linear models. These formulas can be of help in accurately making this decision. Their software program PinT (see Chapter 18) can help researchers to settle these problems. This program calculates standard errors for estimating fixed effects in two-level designs, as a function of sample sizes. The greatest practical difficulty in using this program is that the user has to specify the means, variances, and covariances of all explanatory variables as well as the variances and covariances of all random effects. This specification has to be based on prior knowledge and/or reasonable guesswork. If one is uncertain about these input values, one may try several combinations of them to see how sensitive the resulting optimal sample sizes are for the input values within the likely range.

We refrain from presenting the exact formulas, but give two examples that can help in gaining some notion of how to decide in such a situation. More information can be found in the PinT manual and in Snijders and Bosker (1993).

Example 11.3 A design to study level-one associations.

Suppose one wishes to assess the association between income (INCOME) and total number of years spent in school (YEARS) as part of a larger national survey using face-to-face interviews. Since interviewers have to travel to respondents it seems worthwhile to reduce traveling costs and to take a two-stage sample: randomly select neighborhoods and within each neighborhood select a number of respondents.

First we have to make an educated guess as to the possible structure of the data. The dependent variable is INCOME, while YEARS is the explanatory variable. For convenience let us assume that both variables are standard normal scores (z -scores, with mean 0 and variance 1). (Perhaps monetary income will need to be logarithmically transformed to obtain an approximately normal distribution.) Suppose that for INCOME the intraclass correlation is 0.50 and for YEARS it is 0.30. Now let us guess that the correlation between both variables is 0.447 (which implies that YEARS accounts for 20% of the variation in INCOME). The residual intraclass correlation for INCOME may be 0.40 (since the between-neighborhood regression of INCOME on YEARS generally will be larger than the within-neighborhood regression: dwellings of educated, rich people tend to cluster together). The standard errors of the effect of YEARS on INCOME for various combinations of the sample sizes n and N , restricted so that the total sample size $M = n \times N$ is at most 3,000, are given in Table 11.1.

Since N and n are integer numbers and M may not exceed 3,000, the total sample size fluctuates slightly. Nevertheless, it is quite clear that in this example it is optimal to take $n = 1$ respondent per neighborhood while maximizing the number of neighborhoods in the sample. If we test some target value and assume an effect size of 0.05 (the degree to which the results may be ‘off target’) and set α at 0.01, for $n = 1$ the power $1 - \beta$ would be 0.92, as follows from applying (11.1):

$$0.01348 = \frac{0.05}{2.33 + z_{1-\beta}} \Leftrightarrow z_{1-\beta} = 1.38 \Leftrightarrow 1 - \beta = 0.92.$$

Using $n = 1$ implies a single-level design, in which it is impossible to estimate the intraclass correlation. If one wishes to estimate the within-neighborhood variance and the intraclass correlation, one needs at least two respondents per neighborhood.

Table 11.1: Standard errors as a function of n and N .

$M = N \times n$	N	n	S.E.
3000	3000	1	0.01348
3000	1500	2	0.01390
3000	1000	3	0.01414
3000	750	4	0.01431
3000	600	5	0.01442
3000	500	6	0.01451
2996	428	7	0.01459
3000	375	8	0.01463
2997	333	9	0.01468
3000	300	10	0.01471
3000	250	12	0.01477
2996	214	14	0.01482
2992	187	16	0.01487
2988	166	18	0.01491
3000	150	20	0.01490
2992	136	22	0.01494
3000	125	24	0.01493
2990	115	26	0.01497
2996	107	28	0.01497
3000	100	30	0.01497

In the example given the researchers clearly had an almost unlimited budget. Of course it is far more expensive to travel to 3,000 neighborhoods and interview one person per neighborhood, than to go to 100 neighborhoods and interview 30 persons per neighborhood. It is therefore common to impose budgetary constraints on the sampling design. One can express the costs of sampling an extra macro-unit (neighborhood, school, hospital) in terms of micro-unit costs.

To sketch the idea in the context of educational research, assume that observation costs are composed of salary, travel costs, and the material required. Assume that the costs of contacting one school and the travel for the visit are \$150. Further assume that the salary and material costs for testing one student are \$5. For example, investigating 25 students at one school costs a total of $\$150 + 25 \times \$5 = \$275$. More generally, this means that the cost function can be taken as $\$150N + \$5Nn = \$5N(n + 30)$.

In many practical cases, for some value of c (in the example, $c = 30$) the cost function is proportional to $N(n + c)$. The number c is the ratio indicating how much more costly it is to sample one extra macro-unit (without changing the overall total number of micro-units sampled), than it is to sample one extra micro-unit within an already sampled macro-unit.

Usually for mail or telephone surveys using a two-stage sampling design there is no efficiency gain in using a two-stage sample as compared to using a simple random sample,

so that $c = 0$. But for research studies in which face-to-face interviewing or supervised testing is required, efficiency gains can be made by using a two-stage sampling design.

So we are searching for sample sizes at micro and macro level that satisfy the inequality

$$N(n + c) \leq K, \quad (11.3)$$

in which K is the available budget, expressed in monetary units equal to the cost of sampling one additional micro-level unit. In the example given, K would be the budget in dollars divided by 5.

Example 11.4 Setting up a class size experiment.

Let us now turn to an example in which a macro-level variable is supposed to have an effect on a micro-level variable, and apply the idea just described on cost restrictions.

We might wish to set up an experiment with class size, in which a class size reduction of 6 pupils per class is compared to a usual class size of 26 in its effect on achievement of young pupils. Let us assume that achievement is measured in z -scores, that is, their variance is 1. If there are additional covariates in the design to control for differences between the experimental and control group we assume that the residual variance, given these covariates, is equal to 1.

Let us suppose that the experiment is set up to detect an effect of 0.20 (or larger) in the population of interest. Transforming the effect size d into a correlation coefficient r using the formula (Rosenthal, 1991, p. 20)

$$d = \frac{2r}{\sqrt{1 - r^2}}$$

results in $r = 0.10$. Stated otherwise, the experimental factor will account for $r^2 = 1$ percent of the variation in the achievement score. The astonishing thing about this example is that most researchers would be inclined to test all pupils within a class and would take the class size of 23 (being the average of 20 and 26) as given. There is, however, no need at all to do so. If the total sample size $M = N \times n$ were predetermined, it would be optimal (from a statistical point of view) to take as many classes as possible with only one pupil per class. It is only because of budgetary constraints that a two-stage sample may be preferred. Let us take $c = 23$ (taking an extra school is 23 times the price of taking one extra student within an already sampled school). Running the software program PinT with a budget constraint $N(n + 23) \leq 1,000$ leads to the results in Table 11.2.

The optimum for the standard error associated with the effect of the treatment appears to be when we take a sample of 9 pupils per school and 31 schools. In that case the standard error of interest reaches its minimum of 0.09485. It will be clear from the outset that this standard error is too large to reach a satisfactory power for this effect size of 0.20. Even with α as high as 0.10, the power will be only 0.42.

The first solution to this problem would be to increase the budget. Applying formula (11.1) shows that to reach a power of 0.80 with $\alpha = 0.10$, the standard error should be made twice as small. This implies a multiplication of the number of schools by $2^2 = 4$, which amounts to $4 \times 31 = 124$ schools in each of which 9 pupils are sampled. To make matters even worse, politicians do not like to spend money on interventions that may not work, and for that reason it may be required to have α as low as 0.01. Applying formula (11.1) shows that the original standard error of 0.09485 should be brought down to one third of its value, implying a required sample size of $3^2 \times 31 = 279$ schools.

If, instead of the optimum value $n = 9$, one used the earlier mentioned value of $n = 23$, the standard error would become $0.10346/0.09485 = 1.09$ times as large, which could be offset by sampling $1.09^2 = 1.19$ as many schools. To reach a power of 0.80 with $\alpha = 0.01$ and $n = 23$, one would need to sample $3^2 \times 1.19 \times 31 = 332$ schools.

Table 11.2: Standard errors in case of budget constraints.

$M = N \times n$	N	n	S.E.
41	41	1	0.15539
80	40	2	0.12145
114	38	3	0.10962
148	37	4	0.10267
175	35	5	0.10000
204	34	6	0.09752
231	33	7	0.09602
256	32	8	0.09520
279	31	9	0.09485
300	30	10	0.09487
319	29	11	0.09518
336	28	12	0.09574
351	27	13	0.09652
378	27	14	0.09567
390	26	15	0.09674
400	25	16	0.09798
425	25	17	0.09738
432	24	18	0.09884
437	23	19	0.10046
460	23	20	0.10000
462	22	21	0.10182
484	22	22	0.10144
483	21	23	0.10346

This example illustrates that when one is interested in effects of macro-level variables on micro-level outcomes it usually is sensible (in view of reducing the standard error of interest) to sample many macro-units with relatively few micro-units. The example given may be regarded as a cluster randomized trial: whole classes of students are randomly assigned to either the experimental (a class size of 20) or the control (a class size of 26) condition. Raudenbush (1997) shows how much the standard error for the experimental effect can be decreased (and thus the power increased) if a covariate is added that is strongly related to the outcome variable. In this case this would imply, in practice, to administer a pretest or intelligence test.

The software program Optimal Design (Spybrook et al., 2009) provides power calculations for a macro-level effect on a micro-level outcome with varying intraclass correlations, significance levels, sizes of clusters, numbers of clusters, and varying degrees of associations between the covariate and the dependent variable. If one is interested in a macro-level variable effect on a micro-level outcome, and the macro-units differ considerably with respect to another macro-level variable, one has the choice between matching and covariance adjustment. Raudenbush et al. (2007) discuss this situation, pointing to the fact that

matching in general leads to a larger loss of degrees of freedom, which in general implies that covariance adjustment results in higher power.

11.4.1 Cross-level interaction effects

Of particular interest in social research are cross-level interaction effects. If one is interested in such an effect, how can one improve the power of the test being used? Should one increase the sample size of the macro-units or increase the sample size of the micro-units? If only the total sample size is fixed (i.e., $c = 0$ in (11.3)), then it usually is optimal to take the sample size for the macro-units as large as possible. This is illustrated in the following example, taken from Snijders and Bosker (1993).

Example 11.5 Design for the estimation of a cross-level effect.

Suppose we wish to assess the effect of some school policy measure to enhance the achievement of students from low socio-economic status families, while taking into account aptitude differences between students. More specifically, we are interested in whether this policy measure reduces differences between students from low and high socio-economic status families. In this case we are interested in the effect of the cross-product variable $\text{POLICY} \times \text{SES}$ on achievement, POLICY being a school-level variable and SES being a student-level variable. Assume that we control for the students' IQ and that SES has a random slope. The postulated model is

$$Y_{ij} = \gamma_{00} + \gamma_{10} \text{SES}_{ij} + \gamma_{20} \text{IQ}_{ij} + \gamma_{01} \text{POLICY}_j \\ + \gamma_{11} \text{SES}_{ij} \times \text{POLICY}_j + U_{0ij} + U_{1ij} \text{SES}_{ij} + R_{ij},$$

where Y_{ij} is the achievement of student i in school j .

The primary objective of the study is to test whether γ_{11} is 0. It is assumed that all observed variables are scaled to have zero mean and unit variance. With respect to the explanatory variables it is assumed that SES is measured as a within-school deviation variable (i.e., all school means are 0), the intraclass correlation of IQ is 0.20, the correlation between IQ and SES is 0.30, and the covariance between the policy variable and the school mean for IQ is -0.13 (corresponding to a correlation of -0.3). With respect to the random effects it is assumed that the random intercept variance is 0.09, the slope variance for SES is 0.0075, and the intercept-slope covariance is -0.01 . Residual variance is assumed to be 0.5. Some further motivation for these values is given in Snijders and Bosker (1993).

Given these parameter values, the PinT program can calculate the standard errors for the fixed coefficients γ . Figure 11.1 presents the standard errors of the regression coefficient associated with the cross-level effect as a function of n (the micro-level sample size), subject to the constraint that total sample size M is not more than 1,000. N is taken as M/n rounded to the nearest lower integer. As n increases, N will decrease (or, sometimes, remain constant).

The graph clearly illustrates that it is optimal to take n as small as possible, and (by implication) N as large as possible, since the clustering of the data will then affect the standard errors the least. This may appear to be somewhat counter-intuitive: in the two-step formulation of the model, at the school level we regress the SES slope on the POLICY variable, which suggests that we need a reliable estimate for the SES slope for each school, and that 'thus' n should be large. In this case, however, one should realize that the cross-product variable $\text{POLICY} \times \text{SES}$ is simply a student-level variable, and that as long as we have variation in this variable, the effect can be estimated. This is even when n is as small as 1. Although to estimate intercept and slope variances we need n to be at least 3.

This conclusion can be generalized to studies into the development over time of a certain characteristic (cf. Chapter 15). A series of cross-sectional measurements for people of different ages, for example, can be used to make assessments on growth differences between the sexes.

S.E.($\hat{\gamma}_{11}$)

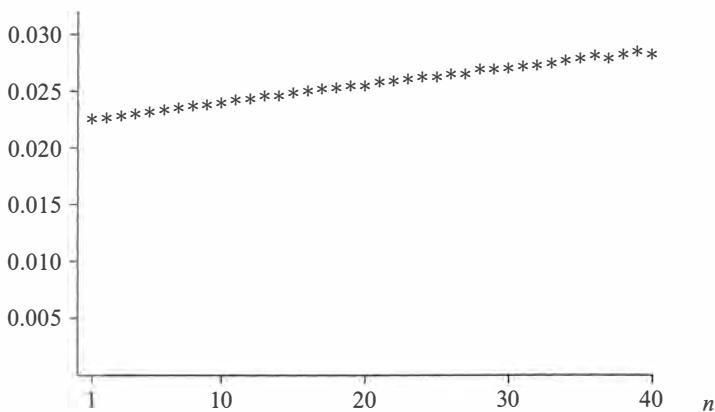


Figure 11.1: Standard error of a cross-level effect (for $N \times n \leq 1,000$).

We again assumed in the previous example that, given the total sample size, it is not extra costly to sample many schools. This assumption is not very realistic, and therefore we will look at the behavior of the relevant standard error once again but now assuming that the price of sampling an extra school is $c = 5$ times the price of sampling a pupil, and the total budget in (11.3) still is $K = 1,000$. Running the software program PinT once again, we obtain the results presented in Figure 11.2.

Because of the budget restriction the total sample size M decreases and the number of groups N increases as the group size n decreases. Balancing between N large (with n small) and N small (with n large), we see that in this case it is optimal to take n somewhere between 15 and 20 (implying that N is between 50 and 40, and M between 750 and 800). More precisely, the minimum standard error is 0.0286, obtained for n equal to 17 and 18. If one is willing to accept a standard error that is 5% higher than this minimum, then n can be as low as 9.

S.E.($\hat{\gamma}_{11}$)

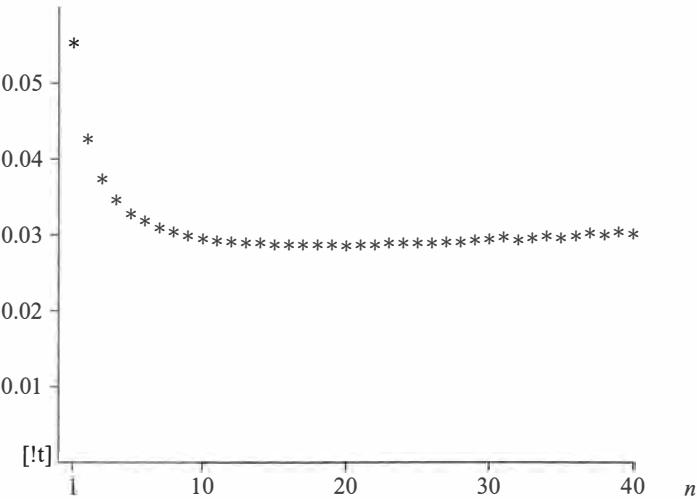


Figure 11.2: Standard error of a cross-level effect (for a budget restriction with $c = 5, K = 1,000$).

11.5 Allocating treatment to groups or individuals

Up to now we have assumed that the design of a multilevel study involves the choice of the number of macro-units and the number of micro-units within those macro-units. In observational designs this is obvious. In many experimental designs, such as the previous example where one is interested in the manipulable macro-variable class size, this is also the case. In some experimental designs, however, one has the choice of manipulating the predictor variable either at the macro level defined by the clusters, or within those clusters. This is known as the difference between cluster randomized trials and multisite trials (Raudenbush and Liu, 2000; Moerbeek, 2005; Moerbeek and Teerenstra, 2011). An example of such a situation is a clinical trial in hospitals, which may be conducted at the level of hospitals (some hospitals giving the experimental treatment to their patients and other hospitals not) or at the level of groups of patients (some receiving the experimental treatment and others not) within hospitals. Or to give an educational example, an experiment where pupils learn words interactively whilst working with a personal computer, which can be delivered at the level of either whole classes or pupils within classes.

Suppose that the research investigates the difference between two experimental treatments. If one determines the treatment at the macro level, the power is determined by the residual intraclass correlation given the treatment administered, $\rho_I(Y|X)$, and the number of micro-level units n_j to be sampled within each macro-level unit j (since these show up in the design effect formula $1 + (n_j - 1)\rho_I$, given in (3.17)). In this case smaller values for n_j and ρ_I lead to improved power, if the total sample size N is kept constant. On the other hand, if one randomizes within the macro-level clusters, then the residual intraclass correlation coefficient has two parts: the similarity of micro-units within the treatment groups within the clusters (ρ_a) and the similarity of micro-units across the treatment groups within the clusters (ρ_b). The latter can be regarded as the consequence of random treatment-by-cluster interaction. The total residual intraclass correlation is $\rho_I = \rho_a + \rho_b$. If for simplicity we assume that the two treatment conditions both are allocated to half of the total sample, the power now is a function of $1 - \rho_a + (n_j - 1)\rho_b$ (Moerbeek, 2005). This shows that a multisite trial is to be preferred over a cluster randomized trial when considering power, because $1 - \rho_a \leq 1$ and $\rho_b \leq \rho_I$.

From a practical point of view, however, multisite trials may be more difficult to implement. For example, one has to make sure that there will be no contamination between the two treatment conditions within the same site. On the other hand, one can argue that a cluster randomized control may lead to selection bias: for example, certain parents might subsequently avoid sending their children to the control schools. A compromise between these two situations is the so called pseudo-cluster randomized trial, where randomization occurs first at the macro level and thereafter at the micro level within the clusters. In this case relatively more micro-units are in the treatment condition within those clusters that were randomly assigned to the treatment condition, whereas the reverse is true for micro-units in the control condition within those clusters that were randomly assigned to the control condition (see Moerbeek and Teerenstra, 2011, for a further discussion on this issue).

11.6 Exploring the variance structure

In all the examples given so far, we have presented cases in which we wished to assess either the mean value of a variable or the association between two variables. Another series of interesting questions may have to do with unspecified effects: does it matter which school a child attends? Do neighborhoods differ in their religious composition? Do children differ in their growth patterns? Is the effect of social status on income different in different regions? In all these instances we are interested in the magnitude of the variance components, and the power of the tests on the variance components is affected by n and N in a different manner from the power of the tests on the fixed effects.

11.6.1 The intraclass correlation

These example questions about unspecified effects can be answered in the first instance by estimating or testing the corresponding intraclass correlation coefficient. Note that in Sections 11.2 and 11.3, the minimum required sample sizes were expressed in terms of the intraclass correlation coefficient, so that also to determine those required sample sizes it is necessary to estimate this coefficient.

In equation (3.13) the standard error for estimating the intraclass correlation was given as

$$S.E.(\hat{\rho}_I) = (1 - \rho_I) (1 + (n - 1) \rho_I) \sqrt{\frac{2}{n(n - 1)(N - 1)}}.$$

To investigate whether it is better to take many small groups or few large ones, a graph of this formula can be drawn as a function of n , the group size, assuming either a constant total sample size or the budget constraint (11.3). For simplicity, let us forget that N must be an integer number, and employ the sample size constraint $M = N \times n$ and the budget constraint $N(n + c) = K$ as equations for computing N as a function of n .

For a fixed total sample size M , the substitution $N = M/n$ implies that the factor with the square root sign in (3.13) must be replaced by

$$\sqrt{\frac{2}{(n - 1)(M - n)}}.$$

The optimal group size is the value of n for which the resulting expression is minimal. This optimum for a fixed total sample size depends on ρ_I . If ρ_I is precisely equal to 0, the standard error decreases with n and the optimum is achieved for $N = 2$ groups each of size $n = M/2$. This implies that to test the null hypothesis $\rho_I = 0$, it is best to use larger group sizes. For ρ_I close to 1, on the other hand, the standard error increases with n and it is best to have $N = M/2$ groups each of size 2. It is unpleasant that the optimum values can be so totally different, depending on the value of ρ_I which is unknown to begin with! However, in most cases the researcher will have some prior knowledge about the likely values of ρ_I . In most social science research, the intraclass correlation ranges between 0.0 and 0.4, and often narrower bounds can be indicated. For such a range of values, the researcher can determine a group size that gives a relatively small standard error for all ρ_I values within this range.

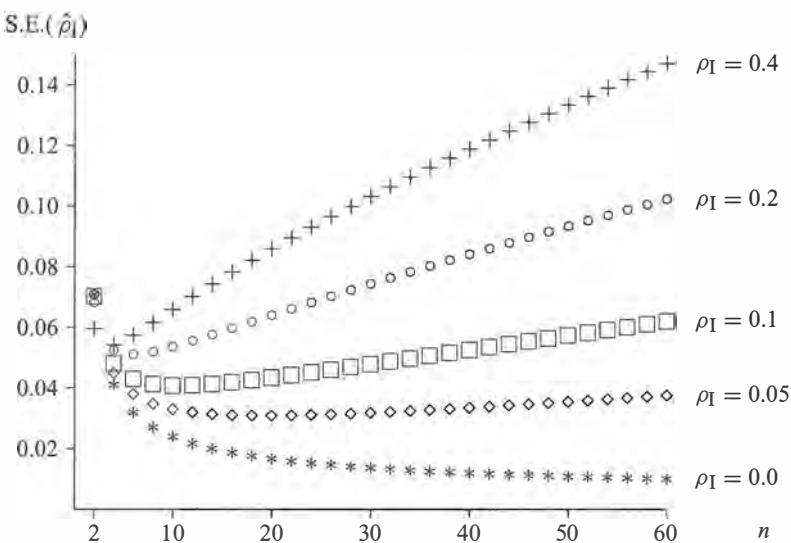


Figure 11.3: Standard error for estimating the intraclass correlation coefficient for $M = 400$.

For a fixed total sample size $M = 400$, the standard error is graphed in Figure 11.3 for several values of ρ_I . This figure shows that the optimum group size does indeed depend strongly on the value of ρ_I . For example, if $\rho_I = 0.4$, group sizes over 20 are unattractive whereas for $\rho_I = 0.0$ this is the case for group sizes of fewer than 5. This suggests that if prior knowledge implies that the intraclass correlation might be as small as 0.0 or as large as 0.4, and if the focus of the research is on estimating ρ_I while the total sample size is fixed at 400, one should preferably use group sizes between 5 and 20.

Example 11.6 *A good compromise for a fixed total sample size.*

Suppose that, as in Figure 11.3, the total sample size must be close to $M = 400$, while it is believed to be likely that the intraclass correlation ρ_I is between 0.05 and 0.2. Calculating the standard errors for these two values of ρ_I shows that for $\rho_I = 0.05$ the smallest standard error is equal to 0.0308, achieved for $n = 19$; for $\rho_I = 0.2$ the minimum is 0.0510, achieved for $n = 6$. This implies that it is best to use an intermediate group size, not too high because that would lead to a loss in standard error for low values of ρ_I , and not too low because that would entail a disadvantage for high ρ_I .

Inspecting the standard errors for intermediate group sizes (this can be done visually from Figure 11.3) shows that group sizes between 9 and 12 are a good compromise in the sense that, for each value $0.05 < \rho_I < 0.2$, the standard error for $9 \leq n \leq 12$ is not more than 10% higher than the minimum possible standard error for this value of ρ_I .

If the optimum group size is to be found for the budget constraint (11.3), the value $N = K/(n + c)$ must be substituted in expression (3.13) for the standard error, and the standard error calculated or graphed for a suitable range of n and some representative values of ρ_I .

Example 11.7 A good compromise for a budget constraint.

Suppose that the budget constraint must be satisfied with $c = 10$, so each additional group costs ten times as much as an additional observation in an already sampled group, and with a total available budget of $K = 600$. The extreme possible situations are then $N = 50$ groups of size $n = 2$, and $N = 2$ groups of size $n = 290$. Suppose that practical constraints imply that the group size cannot be more than 20. The results obtained when substituting $N = 600/(n + 10)$ in (3.13) and calculating the resulting expression are shown in Figure 11.4.

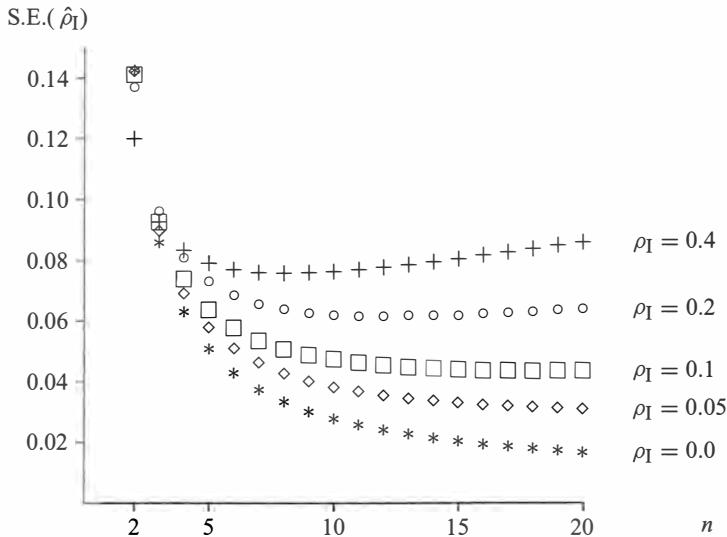


Figure 11.4: Standard error for estimating the intraclass correlation coefficient for a budget constraint with $c = 10$, $K = 600$.

The figure shows that for larger values of ρ_I , the optimum group size has an intermediate value (e.g., $n = 8$ for $\rho_I = 0.4$), but the standard error does not depend very strongly on the group size as long as the group size is at least 4. For the smaller values of ρ_I , it is best to take the group size as large as possible within the allowed range. Group sizes $n = 17$ and 18 give a good compromise between what would be best for ρ_I as small as 0.0 and what would be best for $\rho_I = 0.4$, but smaller group sizes may also be reasonable.

Now suppose, to modify the example, that from the point of view of estimating fixed effects it is desirable to use smaller group sizes, and that it is likely that ρ_I is at least 0.05. On the basis of this figure, one could then decide to sacrifice some of the standard error for low values of ρ_I , and use a group size as low as 8 or 10.

11.6.2 Variance parameters

For a two-level random intercept model, approximations to the estimation variances of the level-one variance σ^2 and the intercept variance τ_0^2 are given by Longford (1993, p. 58). For the standard errors, these formulas are

$$\text{S.E.}(\hat{\sigma}^2) \approx \sigma^2 \sqrt{\frac{2}{N(n-1)}}$$

and

$$\text{S.E.}(\hat{\tau}_0^2) \approx \sigma^2 \frac{2}{Nn} \sqrt{\frac{1}{n-1} + \frac{2\tau_0^2}{\sigma^2} + \frac{n\tau_0^4}{\sigma^4}}.$$

If one prefers standard errors for the estimated standard deviations, according to formula (6.2) these formulas can be transformed by

$$\text{S.E.}(\hat{\sigma}) \approx \frac{\text{S.E.}(\hat{\sigma}^2)}{2\sigma}, \quad (11.4)$$

and similarly for $\text{S.E.}(\hat{\tau}_0)$.

The standard error of $\hat{\sigma}^2$ is minimized by using large level-one sample sizes n , but in practice the standard error is, for a given reasonable budget constraint, so little sensitive to the value of n that one does not really care about obtaining precisely the minimal possible standard error for this parameter (this is also the conclusion reached by Cohen, 1998). To obtain a good design for estimating the variance parameters, the intercept variance is more important.

If optimal sample sizes are to be determined to estimate the intercept variance, the same approach can be applied as in the preceding section. One can plot these standard errors against n , with $M = N \times n$ fixed to a given value, for various values of the intraclass correlation coefficient $\rho_I = \tau_0^2 / (\sigma^2 + \tau_0^2)$. If there is a budget constraint, one can substitute $N = K / (n + c)$ into Longford's approximation expressions and plot the standard errors while keeping $N \times (n + c)$ fixed. An example is given by Cohen (1998, Section 5.1).

Standard errors of random slope variances are quite complicated functions of the sample sizes, the variation of the explanatory variables within and between the groups, and the variances themselves. Further research is needed for elaborating guidelines on how to choose sample sizes appropriately for the estimation of variance parameters in random slope models. If precise estimation of such parameters is required, it seems advisable to use large sample sizes (30 or higher) at either level and to make sure that the explanatory variable of which one wishes to estimate the random slope variance has enough dispersion within the level-two units. This may be achieved, for example, by some kind of stratification.

For complicated models for which no formulas are available for the standard errors, Monte Carlo simulation is an alternative way to form an impression about standard errors. Simulation analysis of multilevel models is possible in various programs, such as MLwiN, MLPowSim (Browne et al., 2009), and the R package pamm (see Chapter 18).

11.7 Glommary

Hypothesis testing. For the general theory of statistical hypothesis testing we refer to the general statistical literature, but in this glommary we mention a few of the basic concepts in order to refresh the reader's memory.

Null hypothesis. A set of parameter values such that the researcher would like to determine whether these are tenable in the light of the currently analyzed data, or whether there is empirical evidence against them; the latter conclusion is expressed as 'rejecting the null hypothesis'.

Alternative hypothesis. A set of parameter values different from the null hypothesis, such that if these values were obtained, the researcher would like the statistical test to conclude that there is evidence against the null hypothesis. In practice, the alternative hypothesis often embodies the theory proposed or investigated by the researcher, and the evidence against the null hypothesis will also be interpreted as support for the alternative hypothesis.

Effect size. Numerical expression for the degree to which a parameter value is believed to deviate from the null hypothesis. This is an ambiguous concept, because the importance of deviations from a null value may be expressed in so many ways. Cohen (1992) gives a list of definitions of effect sizes.

To express differences between groups, the difference between the means divided by the within-group standard deviation is a popular effect size, called Cohen's d . Rosenthal (1991) proposes to express effect sizes generally as correlation coefficients. For multilevel analysis this can be exemplified by noting that variances on their original scales have arbitrary units and therefore are not useful as effect sizes, but the intra-class correlation coefficient transforms the within- and between-group variances to the scale of a correlation coefficient, and thereby becomes comparable across data sets and useful as an effect size index.

Type I error. Mistakenly rejecting the null hypothesis based on the observed data whereas in actual fact the hypothesis is true.

Significance level. Also called type I error rate, and usually denoted α , this is the probability of making a type I error, given parameter values that indeed conform to the null hypothesis. In classical statistical hypothesis testing, the significance level is predetermined by the researcher (e.g., $\alpha = 0.05$; the popularity of this value according to the statistician R.A. Fisher is caused by the fact that humans have 20 fingers and toes). In practice, significance levels are set at lower values accordingly as sample sizes are higher, in view of the tradeoff mentioned below.

Type II error. Failing to reject the null hypothesis when in fact the null hypothesis is false, and the alternative hypothesis is true.

Type II error rate. Often denoted β , this is the probability of making a type II error, given parameter values that indeed conform to the alternative hypothesis. There is a trade-off between the probabilities of making type I and type II errors: having a higher threshold before rejecting the null hypothesis leads to lower type I and higher type II error rates.

Power. This is the probability $1 - \beta$ of correctly rejecting the null hypothesis given parameters that conform to the alternative hypothesis. Power increases as the sample size increases and as effect size increases. But power is also dependent on the type I error rate: a smaller level of significance (smaller type I error rate) will lead to less power (higher type II error rate).

Design effect of a two-stage sample. The ratio of the variance of estimation obtained with the given sampling design, to the variance of estimation obtained for a simple random sample from the same population, and with the same total sample size.

Required sample sizes for two-stage samples, needed to obtain a given standard error of the estimated population mean, and thereby a given power for distinguishing the population mean from a hypothesized value, can be calculated using the design effect.

Power analysis. A study in which one investigates the power of a statistical test to be employed, as a function of the type I error rate α , the sample size, and the effect size. Typically in multilevel studies there are two other considerations: the sample size of the macro-level units of clusters versus the size of the sample of the micro-level units within those clusters, and the residual intraclass correlation, being the degree of resemblance of the micro-level units on the variable of interest, after conditioning on a set of covariates.

Cost function. The cost of the study design in terms of, for example, the sample size. In multilevel situations one weighs the costs of sampling macro-level units against the costs of sampling micro-level units within macro-level units. Such a cost function can then be used to determine optimal sample sizes at either level given a budget constraint.

Cluster randomized trial. An experiment in which whole clusters are randomly assigned to either the intervention or control condition. A typical case is a study in which one assigns schools to conditions and studies the relevant outcomes of the intervention at the level of the pupils. The risk is selection bias if, for instance, certain groups of parents subsequently avoid sending their children to the control schools.

Multisite trial. An experiment in which groups of micro-level units are randomly assigned to the experimental conditions within the clusters. A typical case is a study in which patients within hospitals are assigned to either an experimental treatment or a standard treatment. The risk is contamination: the physician may apply aspects of the experimental treatment to the patients in the control group, without meaning to do so.

Pseudo-cluster randomized trial. A compromise between the cluster randomized trial and the multisite trial to avoid the risk of both contamination and selection bias as much as possible. In this case clusters are randomly assigned to treatment and control conditions, and thereafter micro-level units are randomly assigned to either the experimental or the control condition within their cluster, with a probability depending on whether the cluster is in either the experimental or control condition. In this case a patient may be in the experimental group although the hospital as such was chosen to be in the control group, and vice versa.

12

Other Methods and Models

Several alternative methods for estimation exist, as well as several models closely related to the hierarchical linear model, which can be used in conjunction with, or as alternatives to, the maximum likelihood-based methods presented in the other chapters. Two alternative estimation methods and one alternative model are briefly presented in this chapter.

OVERVIEW OF THE CHAPTER

We begin by introducing the Bayesian approach, with the associated Markov chain Monte Carlo algorithms. This is an alternative estimation method which is more time-consuming but also more flexible, and can be used to estimate more complex models than the maximum likelihood procedures. A global overview is given, with pointers to the literature.

Then we mention the so-called sandwich estimators for standard errors, also called robust standard errors, because they give some degree of robustness to deviations from the model assumptions. Of interest here are the cluster-robust standard errors used in the approach of generalized estimating equations, where parameters are estimated according to a traditional single-level technique and combined with robust standard errors that control for the clustered nature of the data.

We conclude the chapter by a brief treatment of latent class models, which can be used as multilevel models in which the random effects do not have normal distributions but discrete distributions. If the discrete distribution for the random effects has a lot of values, then such a distribution can approximate any distribution. Therefore this can be regarded as a nonparametric or semi-parametric approach, being a combination of a linear or generalized linear model for the mean structure and a nonparametric model for the distribution of the random effects.

12.1 Bayesian inference

The Bayesian approach to statistics is based on a different interpretation of probability than the more usual frequentist approach. Probability in a Bayesian sense expresses the belief in how likely it is that an event will occur or that a variable has a value in a certain

range, a belief which can range from generally shared to subjective. In the frequentist approach, probability is the relative frequency of occurrence of an event in a hypothetical large number of independent replications; this approach is the unmentioned sea in which we are swimming in the rest of this book.

Since the 1980s the Bayesian approach has received increased attention because of the development of computer-intensive so-called *Markov chain Monte Carlo* (MCMC) methods which allow very flexible modeling in the Bayesian paradigm. This has made Bayesian statistical procedures available for quite complicated statistical models, and for models that are hard to manage by frequentist procedures such as maximum likelihood (ML) estimation. An additional advantage of the Bayesian approach is how it handles what statisticians sometimes call *error propagation*, the consequences of uncertainty in various parameters, or aspects of model specification, for the inferential conclusions drawn. The philosophical sharpness of the contrast between the Bayesian and frequentist inferential paradigms has subsided because of a convergence of interpretations and procedures. One reason is that Bayesian procedures often have good frequentist properties for medium and large sample sizes. Another reason is the possibility, in various models, of interpreting probabilities in a frequentist as well as a Bayesian way; for example, the probability distribution of random effects in multilevel models. Treatments of all these issues can be found in recent good textbooks on Bayesian statistics, for example, Gelman et al. (2004) and Jackman (2009).

Outline of Bayesian statistics

A basic element in Bayesian statistics is that the parameters of the statistical model, of which the values are unknown, have a probability distribution, expressing the beliefs and uncertainty of the researcher as to their likely values. The probability distribution of the parameters, entertained by the researcher before having seen the data, is called the *prior distribution*. Observing the data, and applying the statistical procedures, transforms this into the *posterior distribution*. This transformation is effectuated formally through Bayes' rule, which is a direct application of the rules of conditional probability. A major question for Bayesian inference is the choice of the prior distribution. If the researcher has good prior knowledge about the parameters, then this can be reflected by a corresponding choice of the prior. If there is no such prior knowledge, however, it is desirable that the inferential results are not sensitive to the choice of the prior distribution within a reasonable range. The choice of adequate priors is a matter of ongoing debate and research.

The basic correspondence between the posterior distribution and frequentist inferential procedures is as follows:

- As a *point estimate*, the mean or the mode of the posterior distribution can be used: the *posterior mean* or *posterior mode*.
- As a *standard error*, the standard deviation of the posterior distribution can be used: the *posterior standard deviation*.
- A *confidence interval* of confidence level $1 - \alpha$ can be obtained as an interval of parameter values that has probability $1 - \alpha$ under the posterior distribution.
- A *hypothesis test* can be obtained by rejecting the null hypothesis if the parameter value defining the null hypothesis is not included in the confidence interval so

obtained; and the p -value is the smallest value of α such that the null parameter value is not included in the confidence interval of confidence level $1 - \alpha$.

An important result underlying the convergence between Bayesian and frequentist statistics is that these transformations of Bayesian to frequentist procedures have approximately the desired frequentist properties (e.g., correct type I error rates of hypothesis tests) for medium and large sample sizes (Gelman et al., 2004, Chapter 4).

Bayesian MCMC procedures are simulation-based algorithms producing values that can be used as a sample, that is, a sequence of random draws, from the posterior distribution. An important issue here is the convergence of the algorithm, and the dependence between consecutively produced values. Convergence is an issue because the algorithm produces a sequence of values that *converges* to the posterior distribution; for the usual procedures at any given moment there is no full certainty that the algorithm has come close enough to convergence, and convergence diagnostics must be used to check this. Dependence is an issue because the consecutive values are a Markov chain and will be dependent, therefore a larger sample of draws from the posterior distribution will be required than would be the case if the draws were independent. The checks and requirements for coping with these issues depend on the implementation of the algorithm. More and more computer programs for multilevel analysis have incorporated Bayesian MCMC procedures, and these issues are treated in the manuals.

There is a similarity between the Bayesian approach and the hierarchical linear model, which already appeared in the use of empirical Bayes methods in Chapter 4. This can be explained as follows. The Bayesian model for statistics has a hierarchy where first the parameter is drawn from the prior, and then, for this parameter value, the data are drawn from the data-generating distribution. This can be symbolized by

$$\begin{aligned}\theta &\sim \text{prior distribution;} \\ (Y | \theta) &\sim P_\theta,\end{aligned}$$

where θ is the parameter, Y is the data, the vertical bar $|$ denotes conditioning, and P_θ is the probability distribution of the data given parameter value θ . The hierarchical linear model has the same type of hierarchy: in a frequentist setup, the two-level model can be symbolized as

$$\begin{aligned}U &\sim Q_\tau; \\ (Y | U) &\sim P_\theta(U),\end{aligned}$$

where U is the vector of all level-two random effects, Q_τ its distribution with parameter τ , and $P_\theta(U)$ is the probability distribution of the data, conditional on the random effects U , and given parameter θ . In a Bayesian approach, the two hierarchies can be nicely superimposed:

$$\begin{aligned}(\tau, \theta) &\sim \text{prior distribution;} \\ (U | \tau) &\sim Q_\tau; \\ (Y | (U, \theta)) &\sim P_\theta(U).\end{aligned}$$

Here τ contains the parameters of the level-two model, and θ those of the level-one model.

Browne and Draper (2000) presented Bayesian MCMC methods for linear and nonlinear multilevel models; see also further publications by these authors, presented in Draper (2008). In extensive simulation studies, Browne and Draper (2006) found confirmation for the good frequentist properties of Bayesian methods for multilevel modeling. Concise introductions to Bayesian methods for multilevel modeling are given in Raudenbush and Bryk (2002, Chapter 13), Gelman et al. (2004, Chapter 15), Gelman and Hill (2007, Chapter 18), and Hamaker and Klugkist (2011). More extensive treatments are Draper (2008) and the textbook by Congdon (2010).

For the regular (linear) hierarchical linear model, Bayesian and ML procedures yield very similar results. Jang et al. (2007) present an extensive comparative example. An advantage of the Bayesian approach may be the more detailed insight into the likely values of parameters of the random part of the model, because of the deviation from normality of their posterior distributions; in frequentist terms, this is reflected by nonnormal distributions of the estimators and by log-likelihoods, or profile log-likelihoods, that are not approximately quadratic. For more complicated models (e.g., with crossed random effects or with nonnormal distributions), Bayesian approaches may be feasible in cases where ML estimation is difficult, and Bayesian approaches may have better properties than the approximations to some of the likelihood-based methods proposed as frequentist methods. Examples of more complicated models fitted by a Bayesian approach are given in Chapter 13 and in Browne et al. (2007).

12.2 Sandwich estimators for standard errors

For many statistical models it is feasible and sensible to use ML estimators or their relatives (e.g., restricted maximum likelihood (REML) estimators). Statistical theory has a standard way of assessing the precision of ML estimators if the model is correct. This uses the so-called Fisher information matrix, which is a measure for how strongly the probabilities of the outcomes change when the parameters change. The key result is that the large-sample covariance matrix of the ML estimator is the inverse of the information matrix. Recall that standard errors are contained in the covariance matrix of the estimator, as they are the square roots of its diagonal elements. For the hierarchical linear model, this is discussed in de Leeuw and Meijer (2008b, pp. 38–39).

Sometimes, however, the researcher works with a misspecified model (i.e. one that is not a good approximation of reality), often because it is easier to do so; sometimes also the estimation was done by a method different from ML. In such cases, other means must be utilized to obtain standard errors or the covariance matrix of the estimator. A quite general method is the one affectionately called the sandwich method because the mathematical expression for the covariance matrix features a matrix measuring the variability of the residuals, sandwiched between two matrices measuring the sensitivity of the parameter estimates to the observations as far as these are being used for obtaining the estimates. The method is explained in many places, for example, de Leeuw and Meijer (2008b, Section 1.6.3) and Raudenbush and Bryk (2002). It was proposed by White (1980), building on Eicker (1963) and Huber (1967), with the purpose of obtaining standard errors for ordinary least squares (OLS) estimates valid in the case that the linear model holds, but the assumption of independent homoscedastic (constant-variance) residuals is not valid. Therefore he called them *robust standard errors*.

One of the applications is to multilevel structures. There the sandwich estimator yields *cluster-robust standard errors* (Zeger et al., 1988). The term *clusters* here refers to the highest-level units in a multilevel data structure. Thus, even for a multilevel data structure where a linear model is postulated and one suspects within-cluster correlations, one could estimate the parameters of the linear model by ordinary least squares, which is the ML estimator for the case of independent homoscedastic residuals; and then obtain the standard errors from an appropriate sandwich estimator to ‘correct’ for the multilevel structure in the data. This can also be done for three- and higher-level structures, where the clusters are defined as the highest-level units, because these are the independent parts of the data set.

This has been elaborated in the technique of *generalized estimating equations* (GEE), proposed by Liang and Zeger (1986) and Zeger et al. (1988); see the textbook by Diggle et al. (2002). Formulated briefly, this method assumes a linear or generalized linear model for the *expected values* of the dependent variable in a multilevel data structure, conditional on the explanatory variables; but no assumptions are made with respect to the variances and correlations, except that these are independent between the clusters (the highest-level units). The parameters of the linear model are estimated under a so-called ‘working model’, which does incorporate assumptions about the variances and correlations; in the simplest case, the working model assumes uncorrelated residuals with constant variance. The standard errors then are estimated by a sandwich estimator. For large numbers of clusters, this will be approximately correct, provided that the linear model for the means is correct and the highest-level units are independent. The comparison between GEE modeling and the hierarchical linear model is discussed by Gardiner et al. (2009).

Thus, the sandwich method for obtaining standard errors is a large-sample method for use when the model used for estimation is misspecified, or when a different estimator than the ML or REML estimator is used, as in the methods explained in Section 14.5 taking account of design weights in surveys. The sandwich estimator will be less efficient than the ML or REML estimator if the model is correctly specified, and a practical question is for which sample sizes it may be assumed to work in practice. This will also depend on the type of misspecification, and here we should distinguish between misspecification of the correlation structure, such as GEE estimation of a data set satisfying a random slope model with a working model of independent residuals, and misspecification of the distributional form of the residuals, for which the default assumption of normality might be violated.

The quality of cluster-robust standard errors is better when the clusters have equal sizes and the same distributions of explanatory variables, and suffers when the clusters are very different. Bell and McCaffrey (2002) proved that the cluster-robust sandwich estimator for the GEE with the OLS working model is unbiased if all clusters have the same design matrix, and otherwise has the tendency to underestimate variances. Corrections to the sandwich estimator for standard errors were proposed by Mancl and DeRouen (2001), Bell and McCaffrey (2002), and Pan and Wall (2002). These corrections lead to clear improvements of the type I error rates of tests based on cluster-robust standard errors, particularly for clusters with unequal sizes or different distributions of explanatory variables. These corrections should be implemented more widely as options, or even defaults, in software implementations of cluster-robust standard errors. The minimum correction one should apply in the case of a small number of clusters is the following very simple correction mentioned by

Mancl and DeRouen (2001): multiply the estimated covariance matrix as defined by White (1980) by $N/(N - q - 1)$ (this is the HC1 correction of McKinnon and White, 1985), and test single parameters against a t distribution with $N - q - 1$ degrees of freedom (rather than a standard normal distribution), and multidimensional parameters against an F distribution with $N - q - 1$ degrees of freedom in the denominator (rather than a chi-squared distribution). Here N is the number of clusters and q the number of explanatory variables at the highest level. In a simulation study for two-level binary data, Mancl and DeRouen (2001) found that this gives satisfactory results for $N = 40$, and in the case of equal cluster sizes and identical distributions of explanatory variables per cluster, for considerably smaller N . But it is preferable to use one of the more profound and more effective corrections.

Verbeke and Lesaffre (1997) proved that for a hierarchical linear model in which the random part of the model is correctly specified but has nonnormal distributions, an appropriate sandwich-type estimator is consistent, and found support for its performance for moderate and large sample sizes. Yuan and Bentler (2002) also derived sandwich estimators for the hierarchical linear model, together with rescaled likelihood ratio tests for use under nonnormality of residuals. Maas and Hox (2004) reported a simulation study designed to assess the robustness of the sandwich standard error estimator for nonnormal distributions of the level-two random coefficients. They compared the ML estimator and the sandwich estimator, as implemented in MLwiN, for linear models where the covariance structure is correctly specified but the distributions of the level-two random effects are nonnormal. For the fixed effects they found good coverage rates of the confidence intervals based on either standard error for a small number (30) of clusters. For confidence intervals for the level-two variance parameters they found that the sandwich estimator led to better coverage rates than the ML estimator, but even for 100 clusters the coverage rates based on the sandwich estimator were still far from satisfactory for some of the nonnormal distributions. However, Verbeke and Lesaffre (1997) reported better results for a differently defined sandwich estimator under similar distributional assumptions. It can be concluded that the robustness of the sandwich estimator for inference about variance parameters depends on the number of clusters being sufficiently large, or on the appropriate use of specific small-sample versions.

Next to the sandwich estimators, other estimators for standard errors also are being studied in the hope of obtaining robustness against deviations from the assumed model. Resampling methods (van der Leeden et al., 2008) may be helpful here. Cameron et al. (2008) proposed a bootstrap procedure that offers further promise for cluster-robust standard errors.

To decide whether to use an incompletely specified model and the associated sandwich estimator, the following issues may be considered:

1. Can the research questions be answered with the incompletely specified model? In practice, the model will specify the fixed effects and not attempt a correct specification of the variance and correlation structure. In Section 2.2 we argued that the variance and correlation structure, with its parameters such as the intraclass correlation and issues such as heteroscedasticity, may often be of interest, even if they may lie outside the primary scope of the originally conceived aims of the researcher.

An issue in the use of partially specified models is that only those parameters can be meaningfully studied for which there is a correspondence between the true model and the model used for estimation (the ‘working model’ of the GEE approach). Freedman

(2006) goes further, and doubts the wisdom of working with misspecified probability models. One could say that a minimum requirement for the correspondence between parameters in the working model and the true model is consistency and approximate unbiasedness of the parameter estimates. In the example where the true model is a hierarchical linear model as in (5.15) and the working model is the linear regression model with independent homoscedastic residuals, there is such a correspondence for the fixed effects but not for the parameters of the random part. A case where the correspondence does extend to the parameters of the random part is when the covariance structure of the random part is well specified but the distributions are nonnormal, as in Verbeke and Lesaffre (1997) and Maas and Hox (2004).

2. If one uses a completely specified model, is there enough confidence that its assumptions are a reasonable approximation to reality? To answer such questions, diagnostic methods and model improvements as discussed in Chapter 10 can be used.
3. Does the sandwich estimator provide approximately correct standard errors of an approximately unbiased estimator? Here the finite-sample corrections discussed above are relevant. It should be realized that if the model is approximately correct, REML estimators of standard errors will be more precise (have smaller variability) than the sandwich estimator. For 40 or more clusters, the sandwich estimator may be expected usually to have good properties if the HCl correction and finite-sample degrees of freedom are employed, as mentioned above, unless there is great imbalance between the clusters. For fewer clusters, this depends on the design and parameter values, and one may need additional arguments or better corrections to justify the use of the sandwich estimator.
4. Is the loss of efficiency (of parameters and tests) acceptable as compared to the efficiency of using ML methods under an approximately correctly specified model – assuming that it is feasible to obtain such a model? The loss of efficiency will depend on the study design, the imbalance between the clusters, and the parameters of the random part.

Example 12.1 Sandwich estimator for the language test.

Here we take up the model of Example 5.4. In the GEE approach, the parameters of the linear model are estimated by OLS, and the standard errors by the sandwich estimator. For 211 groups, the correction methods mentioned above are small; with $q = 3$ level-two variables, the correction factor of the HCl method is $N/(N - q - 1) = 1.02$.

Table 12.1 presents the ML estimates for the hierarchical linear model alongside the OLS estimates (with REML estimate for the residual variance) and the sandwich estimates for the standard errors. The difference between the OLS estimates and those from the hierarchical linear model seem small, but for the classroom average of IQ the difference of $0.986 - 0.760 = 0.226$ is more than 75% of a standard error, and therefore not really small. The estimates from the hierarchical linear model here may be considered to be more efficient, because they take account of the dependence between individuals in the same class. The sandwich standard errors are very close to the standard errors from the hierarchical linear model.

Table 12.1: Estimates for the hierarchical linear model compared with OLS estimates and sandwich standard errors.

	Hierarchical linear model		Ordinary least squares	
Fixed effect	Coefficient	S.E.	Coefficient	Sandwich S.E.
γ_{00} = Intercept	41.612	0.247	41.686	0.259
γ_{10} = Coefficient of IQ	2.231	0.063	2.207	0.067
γ_{20} = Coefficient of SES	0.174	0.012	0.174	0.012
γ_{30} = Interaction IQ \times SES	-0.017	0.005	-0.017	0.005
γ_{01} = Coefficient of \overline{IQ}	0.760	0.296	0.986	0.287
γ_{02} = Coefficient of \overline{SES}	-0.089	0.042	-0.101	0.041
γ_{03} = Interaction $\overline{IQ} \times \overline{SES}$	-0.120	0.033	-0.108	0.039
Random effect	Var. comp.	S.E.		
<i>Level-two random effects:</i>				
$\tau_0^2 = \text{var}(U_{0j})$	8.369	1.050		
$\tau_1^2 = \text{var}(U_{1j})$	0.164	0.069		
$\tau_{01} = \text{cov}(U_{0j}, U_{1j})$	-0.929	0.204		
<i>Level-one variance:</i>				
$\sigma^2 = \text{var}(R_{ij})$	37.378	0.907	46.04	
Deviance	24,626.8			

12.3 Latent class models

The usual assumption for random coefficients in a hierarchical linear model is that they have normal distributions (Chapter 10); this is sometimes weakened to heavier-tailed distributions such as t distributions (e.g., Seltzer and Choi, 2003) or mixtures of normals (Verbeke and Lesaffre, 1996; Verbeke and Molenberghs, 2000, Chapter 12). Another possibility is to assume that these random coefficients have arbitrary discrete distributions. Consider the two-level hierarchical linear model as defined in (5.15):

$$Y_{ij} = \gamma_0 + \sum_{h=1}^r \gamma_h x_{hij} + U_{0j} + \sum_{h=1}^p U_{hj} x_{hij} + R_{ij}.$$

If the vector of random coefficients $(U_{0j}, U_{1j}, \dots, U_{pj})$ has a discrete distribution with C values u_1, \dots, u_C , then each level-two unit has one of these values, so that the population of all level-two units can be divided into C categories, the so-called *latent classes*, each with a separate value u_c ($c = 1, \dots, C$). This can be called a latent class model or a finite mixture model, because the distribution is a mixture of C basic distributions. Such models were studied by, among others, Wedel and DeSarbo (2002), Vermunt (2004), and Rabe-Hesketh et al. (2005). Good and extensive treatments are given by Vermunt (2008) and Muthén and Asparouhov (2009). A similar kind of modeling can be applied to generalized linear

models where the dependent variable is not normally distributed. Such models are included in the very general framework of Skrondal and Rabe-Hesketh (2004, Chapter 4). They can be fitted by software such as Latent Gold, gllamm, Mplus, and WinBUGS. A theoretical exposition and some examples for discrete and continuous dependent variables, for two and three levels, are given in Vermunt (2011).

If the number C of latent classes is allowed to be large enough, this model yields the nonparametric ML estimator for the hierarchical linear model, that is, the ML estimator for the model where the hierarchical linear model is specified as in (5.15) but without any assumption as to the distribution of the random coefficients $U_{0j}, U_{1j}, \dots, U_{pj}$ (Laird, 1978). ML estimation has a tendency toward overfitting, however, meaning here toward overestimating the number of latent classes. More latent classes will always yield a better fit in the sense of a larger likelihood (i.e., a smaller deviance). Therefore it may be preferable to choose the number of latent classes based on an information criterion such as the Bayesian information criterion (BIC); see, for example, Skrondal and Rabe-Hesketh (2004, Section 8.4). The BIC is defined by

$$\text{BIC} = \text{deviance} + p \ln(M),$$

where p is the number of parameters and $\ln(M)$ is the natural logarithm of the total sample size. It is used by selecting the model for which the BIC is minimal; this gives a tradeoff between good fit (low deviance) and a small number of parameters.

Lukočienė and Vermunt (2007, 2009) compared the model with normally distributed random effects with the latent class model for two-level logistic regression models. They found that for the random intercept model the latent class model performs very well, even for random effects that in reality have a normal distribution. For a model with a random slope it appeared that the latent class model combined with the ML estimator is less satisfactory; the model with normally distributed random effects, as well as the latent class model where the number of classes is estimated by the BIC, both have good properties, and which of these is preferable depends on the true random effects distribution. It should be noted that this is specific to the model for binary outcomes, and does not carry over to models that have, within the latent classes, normally distributed dependent variables.

Example 12.2 Latent class version of the model for the language test.

Again we take up the model of Example 5.4, and now estimate it with a latent class model, that is, a model with the same linear specification but a discrete distribution for the random coefficients (U_{0j}, U_{1j}) . If the number of classes is C , this means that there are C possible values for the pair (U_{0j}, U_{1j}) , each with their own probability. The software ‘Latent Gold’ was used.

The estimation for these models has the problem of local optima: the estimation algorithm finds a solution which is not sure to be the best (the best is the global optimum). Therefore the algorithm is executed several times with independent starting values, and one hopes that the best solution is indeed found.

Table 12.2 gives the deviance, the number of parameters p , and the BIC, as a function of the number of latent classes C ; this data set has a total sample size of $M = 3,758$. The deviance becomes smaller and smaller as C increases, which is usual. The BIC favors a model with $C = 3$ latent classes, a very low number. By way of comparison, the deviance of the model with normally distributed random coefficients is 24,626.8 with $p = 11$ parameters (see Table 5.4), yielding $\text{BIC} = 24,717.3$, considerably smaller than for any of the latent class models. This shows that for this data set, the model with normally distributed random coefficients is somewhat more satisfactory than a latent

Table 12.2: Deviance and BIC values.

C	2	3	4	5	6	7	8
Deviance	24,722.8	24,642.8	24,624.9	24,618.2	24,614.8	24,611.6	24,609.9
p	11	14	17	20	23	26	29

BIC	24,813.3	24,758.0	24,764.8	24,782.8	24,804.1	24,825.6	24,848.6
-----	----------	----------	----------	----------	----------	----------	----------

class model. The parameter estimates for the fixed effects for latent class models with $C \geq 4$ are close to the parameter estimates in Table 5.4.

12.4 Glommary

Bayesian statistics. An approach to statistical inference based on the notion that probability is a measure of how likely one thinks it is that a given event will occur or that a variable has a value in a given range. In particular, beliefs and uncertainties about parameters in a statistical model are expressed by assuming that these parameters have a probability distribution.

Frequentist statistics. The more usual approach to statistical inference based on the notion that probability is the relative frequency of occurrence of an event in a hypothetical large number of independent replications. Parameters are considered just as unknown numbers or vectors.

Prior distribution. The probability distribution of the parameters, entertained by the researcher before having seen the data.

Posterior distribution. The probability distribution of the parameters, entertained by the researcher after having seen and analyzed the data.

Posterior mean. The expected value of the posterior distribution, used as a point estimate of the parameters.

Posterior standard deviation. The standard deviation of the posterior distribution, used as a measure of posterior uncertainty about the parameters.

Markov chain Monte Carlo (MCMC) methods. Simulation-based algorithms that produce values that can be used as a sample, that is, a sequence of random draws, from the posterior distribution.

Bayesian procedures for multilevel modeling. These are described in various articles and specialized textbooks.

Sandwich estimator. A special type of estimator for standard errors, or more generally of the covariance matrix of an estimator. This estimator generally is more *robust* against deviation from assumptions concerning the random part of the model. A special case is the *cluster-robust* standard error estimator, which is designed to keep account of the multilevel dependencies (also referred to as dependencies due to clustering) even

if parameters were estimated by a single-level method. For large numbers of units at the highest level, this standard error estimator performs quite well.

Generalized estimating equations (GEE). An estimation method assuming a linear or generalized linear model for the expected values of the dependent variable in a multi-level data structure, conditional on the explanatory variables, but making no specific assumptions with respect to the variances and correlations.

Multilevel latent class models. Variations of the hierarchical linear model, where the random effects have a discrete rather than a normal distribution. This can be regarded as a nonparametric or semi-parametric approach, because discrete distributions with a sufficient number of categories can approximate any distribution. Another name for latent class models is *mixture models*.

13

Imperfect Hierarchies

Up to now we have treated multilevel data structures, and statistical models to handle these, in which a lower-level unit (e.g., a pupil) is perfectly nested in one, and only one, higher-level unit (e.g., a school). Reality, of course, is often far more complicated. Pupils not only belong to schools but also to neighborhoods, and one may be interested in assessing the relative effects on pupils' citizenship behavior in both social contexts simultaneously. Studying interpersonal teacher behavior implies asking pupils how they rate their teacher on several dimensions. But classes of pupils are taught by more than one teacher, and teachers teach more than just one class. Studies about effects of secondary schools on pupil achievement are based on the idea of 'value added', that is, the gain a pupil has made during a certain period (say, five years). But pupils sometimes switch secondary schools during their school career, so more than one secondary school is related to their school success. The aforementioned situations are referred to as *imperfect hierarchies*. Specific models have been developed as special cases of the hierarchical linear model to deal with these structures. They are called *cross-classified models*, *multiple membership models*, and *multiple membership multiple classification models*. In models for cross-classified data a lower-level unit belongs uniquely to one higher-level unit of the first type (e.g., a school) and also uniquely to one higher-level unit of the second type (e.g., a neighborhood), but the two types of unit are not nested either way. In models for multiple membership data the higher-level units are all of one type (e.g., schools), but the lower-level units (pupils in this example) may belong to more than one higher-level unit. In *multiple membership multiple classification models* we have both situations at the same time. In this chapter we will introduce these models for imperfect hierarchies.

OVERVIEW OF THE CHAPTER

In this chapter we first treat cross-classified data structures, implying that a model with crossed random factors is needed to deal with it. After introducing a two-level model with crossed random intercepts, a three-level model with cross-classifications is treated. Multiple membership models are then explained, and some special attention is given to how to weight the duration of membership of a lower-level unit in a higher-level unit. Having presented and clarified the models for cross-classifications and multiple membership, we

then introduce the straightforward combination of these: the multiple membership multiple classification model. These models may become quite complex, and therefore brief mention is made of the Bayesian approach and the associated Markov chain Monte Carlo estimation algorithm, which was treated in Section 12.1.

13.1 A two-level model with a crossed random factor

Consider the case of cross-classified data, where individuals belong to different social units (e.g., schools, neighborhoods, families, peer groups, sports clubs, companies) at the same time. One can think of this situation as crossed factors, that is, individuals are nested within companies but also within neighborhoods, and these two factors are crossed. Each individual uniquely belongs to one combination of both: a specific company as well as a specific neighborhood. These two factors are not nested, as not all employees of a given company live in the same neighborhood, nor do employed people living in a certain neighborhood all work in the same company. One can think of this situation as lower-level units being nested within two factors that cross, or as a two-level data structure (that handles one of the levels of nesting) with another crossed factor (representing the other level of nesting).

To understand the situation a classification graph as suggested by Rasbash and Browne (2001, 2008) may be helpful, as presented in Figure 13.1.

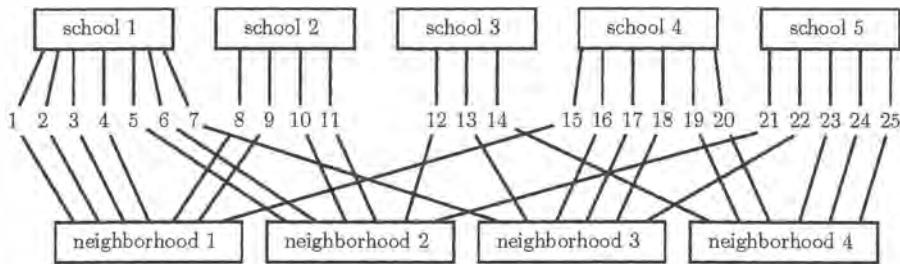


Figure 13.1: An example of pupils nested within a cross-classification of schools and neighborhoods.

In this figure, 25 pupils are nested within five schools, but also within four different neighborhoods. The lines indicate to which school and neighborhood a pupil belongs. The seven pupils of school 1 live in three different neighborhoods. The six pupils who live in neighborhood 3 attend four different schools. Another way to present the cross-classification is by making a cross-table, which of course readers can do themselves.

Following up on this example, we now continue with an explanation of how to incorporate a crossed random factor in a multilevel model. We consider the case of a two-level study, for example, of pupils (indicated by i) nested within schools (indicated by j), with a crossed random factor, for example, the neighborhood in which the pupil lives. The

neighborhoods are indexed by the letter k , running from 1 to K (the total number of neighborhoods in the data). The neighborhood of pupil i in school j is indicated by $k(i,j)$. The hierarchical linear model for an outcome variable Y_{ij} without the neighborhood effect is given by

$$Y_{ij} = \gamma_0 + \sum_{h=1}^q \gamma_h x_{hij} + U_{0j} + \sum_{h=1}^p U_{hj} x_{hij} + R_{ij}$$

(cf. (5.15)). The random effect of neighborhood k , indicated by W_{0k} , is simply added to this model:

$$Y_{i(j,k)} = \gamma_0 + \sum_{h=1}^q \gamma_h x_{hij} + U_{0j} + \sum_{h=1}^p U_{hj} x_{hij} + W_{0k} + R_{ij}. \quad (13.1)$$

There is some ambiguity concerning the second level, as is represented by the notation $Y_{i(j,k)}$. There are two indices at the second level, j as well as k , denoting the nesting in school j and simultaneously in neighborhood k . The nesting in schools is represented by the random effects U_{hj} like in the preceding chapters. For the neighborhood effects W_{0k} , the usual assumption is made that W_{01}, \dots, W_{0K} are mutually independent random variables, also independent of the other random effects (U_{0j} and R_{ij}), W_{0k} having the normal distribution with mean 0 and variance τ_W^2 . The interpretation is just like that of other random effects: part of the variability in the dependent variable is accounted for by neighborhoods, and these are regarded as a random sample from a (perhaps hypothetical) population of neighborhoods. The next subsection indicates how to solve the statistical complications caused by the fact that the structure is not neatly nested.

The analysis by multilevel software of random effects models that combine nested and crossed effects is technically and computationally more demanding than the usual multi-level models. Therefore, it is advisable to make preliminary analyses focusing on first one and then the other crossed factor; such analyses can be done employing the usual two-level model. An important question is to what extent the two crossed factors are associated. In a study of schools and neighborhoods, these two factors will be associated, and when only one of them is incorporated as a random effect in the model it will draw to itself part of the other (temporarily neglected) random effect. In a generalizability study, on the other hand, usually the various crossed factors are not associated and they will have (nearly) independent effects.

This model can be estimated in a frequentist framework as described in Raudenbush (1993) and Rasbash and Goldstein (1994). The software packages HLM, MLwiN, R (packages lme4 and nlme), SAS, and Stata offer ways to do these computations. It turns out, however, that the more complex models can often more easily be estimated within a Bayesian framework, applying the Markov chain Monte Carlo algorithm as discussed in Section 12.1. This is implemented in MLwiN as well (Browne, 2009).

Example 13.1 Sustained primary school effects.

This example is derived from a study on disentangling complex school effects (Timmermans et al., forthcoming), using data from a large-scale Dutch cohort study (Kuyper and Van der Werf, 2003). Around 20,000 pupils in approximately 100 schools were followed after having entered secondary

Table 13.1: Models with and without cross-classification for examination grades.

	Model 1		Model 2	
Fixed effect	Coeff.	S.E.	Coeff.	S.E.
γ_0 Intercept	6.36	0.03	6.36	0.03
Random part	Var. comp.	S.E.	Var. comp.	S.E.
<i>Crossed random effect:</i>				
$\tau_W^2 = \text{var}(W_f)$ primary school			0.006	0.005
<i>Level-two random effect:</i>				
$\tau_0^2 = \text{var}(U_{0j})$ secondary school	0.067	0.014	0.066	0.014
<i>Level-one variance:</i>				
$\sigma^2 = \text{var}(R_{ij})$	0.400	0.010	0.395	0.010

education in 1999 up to the point where they did their final national examinations in a range of subjects, in most cases six or seven. The dependent variable is the average examination score across those subjects, which in theory can range from 1 (low) to 10 (high), but empirically ranges from 3 to 9. A subset of this cohort is taken, namely those pupils who took their final national examination in the lower general education track. This subsample consists of 3,658 pupils in 185 secondary schools, the number being larger than the original 100 because of pupil mobility. It is also known which primary schools these students attended before they entered secondary education. They stem from 1,292 different primary schools. In order to find out whether the effect of the primary school has a sustained effect on the examination score of the pupils at the end of the secondary stage, a cross-classified multilevel model analysis is performed on the data. The nesting structure is pupils within the secondary schools in which they did their final examination, crossed with the primary schools they stem from. Table 13.1 contains the results of a model without (Model 1) and a model with the inclusion of primary school effects (Model 2).¹

In Model 1 in Table 13.1, the average examination score is 6.36 and the total variance 0.466 (difference due to rounding). Of this variance, 14% is associated with the secondary schools. Model 2 presents the results in which the available information on the primary schools previously attended by the pupils is also taken into account. Of course the average examination grade remains the same, but now we see some changes in the variance components. The variance between secondary schools marginally decreases to 0.066, and the within-school variance decreases somewhat as well, since now the primary schools take up a variance component of 0.006. The decrease in deviance is $7025.7 - 6977.4 = 48.3$, highly significant when compared to a chi-squared distribution with $df = 1$. We observe here an instance where, for a variance component, the estimate divided by the standard error is not very large, but yet the variance is significantly different from 0 (cf. Section 6.1).

¹The cross-classified models in this chapter are estimated using the Bayesian MCMC algorithm of Browne (2004) with 10,000 iterations, and parameter expansion at the level of the primary school. For Bayesian procedures, see Section 12.1.

There are now three different intra-class correlations: the correlation between examination grades of pupils who attended the same primary school but went to a different secondary school is

$$\frac{\tau_W^2}{\tau_W^2 + \tau_0^2 + \sigma^2} = \frac{0.006}{0.467} = 0.013;$$

the correlation between grades of pupils who attended the same secondary school but came from different primary schools is

$$\frac{\tau_0^2}{\tau_W^2 + \tau_0^2 + \sigma^2} = \frac{0.066}{0.467} = 0.141;$$

and the correlation between grades of pupils who attended both the same primary and the same secondary school is

$$\frac{\tau_W^2 + \tau_0^2}{\tau_W^2 + \tau_0^2 + \sigma^2} = \frac{0.072}{0.467} = 0.154.$$

We can elaborate the models by including predictor variables. The original data set contained many potential candidate predictors, such as IQ, gender, and the socio-economic status of the pupil's family. For almost all of these variables scores were missing for some pupils, and therefore we employed the chained equations technique described in Chapter 9 to impute values, using all the information available.

From the potential predictors we selected an entry test (grand mean centered) composed of arithmetic, language and information processing subscales; socio-economic status (SES, grand mean centered); the primary school teacher's advice about the most suitable level of secondary education (grand mean centered); and ethnicity. Table 13.2 contains the results with a model including these four predictor variables (Model 3) as an extension to the previously fitted Model 2 (presented again in this table). According to the principles of multiple imputation (Section 9.4) we constructed 25 data sets with imputed values. The results in the table, as well as those in the following tables in this chapter, are the syntheses of 25 analyses run on these imputed data sets, using equations (9.3)–(9.6).

The four predictor variables all have highly significant effects, indicating that pupils with higher entry test scores, with higher recommendations from their primary school teachers, and from more affluent families have higher average examination scores. Moreover, pupils from ethnic minorities have lower examination results than pupils from the Dutch majority group. Most important, however, are the estimates of the variance components. Comparing Models 2 and 3, all variance components have decreased. The between-pupils within-schools variance decreases from 0.395 to 0.330. The between-secondary-schools variance (0.034, and significant) is almost half its original estimate (0.066), which also turns out to be the case for the between-primary-schools variance: this decreases from 0.006 to 0.003. This indicates that secondary schools appear to have a value-added effect on pupil achievement measured at the final examination, but that primary schools, given the achievement levels attained by pupils at the end of primary education and given their family background, have only a marginally lasting effect as measured four or five years later at the secondary school examinations.

Although for the models in the example the predictor variables only had fixed effects, it is straightforward to include these variables with random slopes at the level of one or both of the crossed random factors as shown in equation (13.1).

Table 13.2: Models with and without cross-classification for examination grades (continued).

	Model 2		Model 3	
Fixed effect	Coeff.	S.E.	Coeff.	S.E.
γ_0 Intercept	6.36	0.03	6.39	0.02
γ_1 Pretest			0.032	0.002
γ_2 SES			0.068	0.011
γ_3 Advice			0.054	0.009
γ_4 Ethnicity			-0.071	0.028
Random part	Var. comp.	S.E.	Var. comp.	S.E.
<i>Crossed random effect:</i>				
$\tau_w^2 = \text{var}(W_f)$ primary school	0.006	0.005	0.003	0.003
<i>Level-two random effect:</i>				
$\tau_0^2 = \text{var}(U_{0j})$ secondary school	0.066	0.014	0.034	0.008
<i>Level-one variance:</i>				
$\sigma^2 = \text{var}(R_{ij})$	0.395	0.010	0.330	0.008
Deviance	6,977.4		6,319.9	

13.2 Crossed random effects in three-level models

A crossed random effect in a three-level model can occur in two ways. An example of the first is pupils nested in classes nested in schools, with neighborhoods as a crossed effect. In this case *the extra factor is crossed with the level-three units*: neighborhoods are crossed with schools.

The second kind of random effect in a three-level model is exemplified by pupils nested in schools nested in towns, again with neighborhoods as crossed effects. Here *the extra factor is crossed with the level-two units and nested in the level-three units*: neighborhoods are crossed with schools and nested in towns.

Such models are rather complex, but fortunately can be handled straightforwardly in a Bayesian framework (Section 12.1) using the MCMC algorithm, as implemented, for example, in MLwiN.

13.3 Multiple membership models

When schools and pupils are of focal interest in a study following pupils over time, pupils may have attended more than one school during the period under investigation. Which school is now ‘responsible’ for a pupil’s school success? The first, the last, the ones in between? In Example 13.1 we chose the last school, but others might be more interested in

the first secondary school these pupils attended, as this school gave them the essential head start. Multiple membership models have been developed to analyze this complex reality, thus avoiding the kind of choices just mentioned. That some pupils have attended more than one school can be modeled explicitly. Once again a classification graph, given in Figure 13.2, may clarify what is meant by multiple membership.

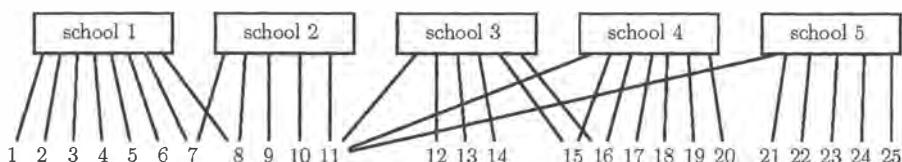


Figure 13.2: An example of pupils nested within multiple schools.

The first six pupils all attended only school 1, but pupil 7 attended both school 1 and school 2. The most extreme case is pupil 11, who attended four different schools. One can think of a situation where the pupil is being bullied and his parents try to resolve the situation by choosing another school. Later they move house, so the pupil once again has to change schools, and exactly the same problem of being bullied occurs, and again a switch to another school is seen as the solution. The main difference with the cross-classified situation, where the individuals belong to units of two different populations of interest at level two (schools and neighborhoods, or primary and secondary schools), is that in the multiple membership situation there is only one such population of interest at level two, but the individuals belong to more than one level-two unit in that population. Other examples are teachers working in more than one school, pupils rating their teachers of different subjects, families who changed neighborhoods during the period under investigation, or employees changing firms.

The solution to the problem is to use some kind of weighting, where the membership weights can be proportional to the time a level-one unit spent at a level-two unit, with the weights summing to 1 (Hill and Goldstein, 1998). So a pupil who attended the same primary school from grade 1 to grade 6 has a membership weight of 1 for that school and 0 for all other schools. A pupil attending the first three grades in her first school and the last three grades in another school has a membership weight of $\frac{1}{2}$ for the first, $\frac{1}{2}$ for the second, and 0 for all other schools. The membership weights are denoted by w_{ih} for pupil i in school h , adding up to 1 for every pupil:

$$\sum_{h=1}^N w_{ih} = 1.$$

The multilevel model for this situation, following the notation of Rasbash and Browne (2001) and ignoring explanatory variables, is:

$$Y_{i\{j\}} = \gamma_0 + \sum_{h=1}^N w_{ih} U_{0h} + R_{i\{j\}}. \quad (13.2)$$

The subscript $\{j\}$ denotes that a level-one unit does not necessarily belong to one unique level-two unit. Therefore the level-two residuals U_{0h} are weighted by w_{ih} . For a given

pupil i , the schools h that this pupil never attended have $w_{ih} = 0$, so they do not contribute anything to the outcome of (13.2). For example, for a pupil who spent a quarter of his time in school 1 and the remainder in school 2, this part of the formula gives the contribution

$$\frac{1}{4} U_{01} + \frac{3}{4} U_{02}.$$

If there are explanatory variables at level two, then a similar kind of weighting is used for these variables.

Next to this contribution from multiple random effects, the division of a pupil's school career over multiple schools could also have a fixed effect, for example, because of the difficulties of adjustment. To represent this as a consequence of the fractionalization of the weights w_{ih} , one may give a fixed effect to the variable

$$W_i = \frac{1}{\sum_h w_{ih}^2} - 1, \quad (13.3)$$

which is 0 if one weight is 1 and all others 0, and positive if two or more of the weights are non-zero; for example, if for pupil i there are K weights of equal values $1/K$ and all other weights are 0, then $W_i = K - 1$.

Example 13.2 Multiple membership in secondary schools.

Following up on the previous example and paying attention only to the secondary school period of a pupil's school career, thus ignoring the sustained effects of the primary schools, we now try to do justice to the fact that many pupils attended more than one secondary school. In this sample 3,438 of the 3,658 pupils did their final examination in the school where they originally enrolled. So 94% of the pupils never changed schools. Of the remaining 6%, 215 pupils attended two, and 5 pupils attended three different schools. For the pupils who attended two different schools the membership weights for their first school vary between 0.20 and 0.80. For the five pupils who attended three different schools the membership weights for their first school vary between 0.20 and 0.25, indicating that they rather quickly left this school. Unlike the analyses in the previous example, we start with a model (Model 4) in which the second level is defined by the first school enrolled, of which there are 86 in total in this subset of the sample. Then a multiple membership model (Model 5) is fitted to the data.

From Table 13.3 it can be seen that the results for Model 4 are very similar to those presented in Model 1 of Table 13.1. The variance for the first school a pupil attended is close to the variance for the last school attended. In Model 5 the results of the multiple membership modeling of the data can be seen. There is a marginal increase in the between-school variance, from 0.062 to 0.064, indicating that the multiple membership model gets somewhat closer to the data than the 'simple' multilevel model. To see whether there is an impact of the more advanced modeling on the estimated school effects, Figure 13.3 shows a scatterplot of the residuals ('school effects') for the school where the pupils did their final examinations.

On the horizontal axis the school-level residuals derived from the 'simple' multilevel (ML) model (Model 4) are plotted. On the vertical axis the school-level residuals derived from the 'more realistic' multiple membership (MM) model are pictured. As can be seen, there is a high correlation (0.86 to be precise) between the two types of residuals, but the most striking point is that the residuals from the multiple membership model are slightly less dispersed. That is, of course, because part of the between-schools variation is now accounted for by the other schools previously attended by some of the pupils.

Table 13.3: Models without and with multiple membership.

	Model 4 (ML)		Model 5 (MM)	
Fixed effect	Coeff.	S.E.	Coeff.	S.E.
γ_0 Intercept	6.36	0.03	6.36	0.03
Random part	Var. comp.	S.E.	Var. comp.	S.E.
<i>Level-two random effect:</i>				
$\tau_0^2 = \text{var}(U_{0j})$ secondary school	0.062	0.013	0.064	0.014
<i>Level-one variance:</i>				
$\sigma^2 = \text{var}(R_{ij})$	0.402	0.010	0.401	0.009

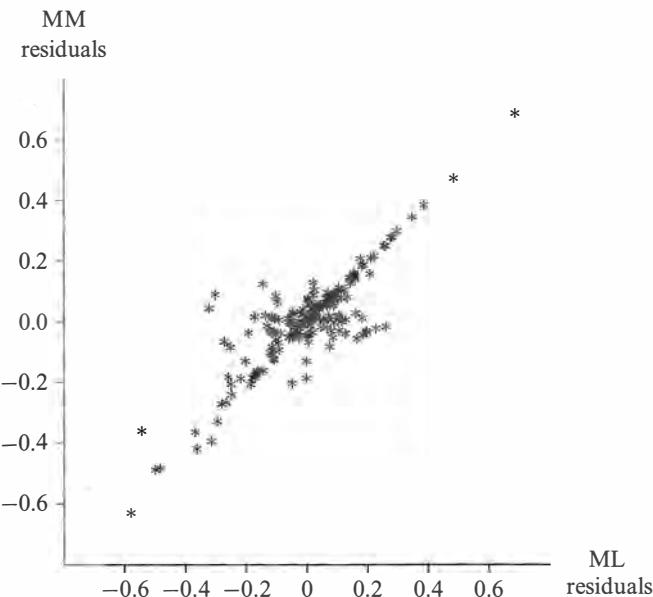


Figure 13.3: Scatterplot of ML- and MM-type school level-two residuals.

13.4 Multiple membership multiple classification models

As may be expected, cross-classified and multiple membership data structures can both be present and of interest to the researcher (Browne et al., 2001). She may, for instance, be interested in the effects of neighborhood as well as of multiple school membership on pupils' attitudes toward ethnic minorities. We can simply combine Figures 13.1 and 13.2 to depict such a situation, yielding the classification graph of Figure 13.4.

The formula to be used for such a multiple membership multiple classification situation consists of a merger of formula (13.1) for a cross-classified model with formula (13.2) for

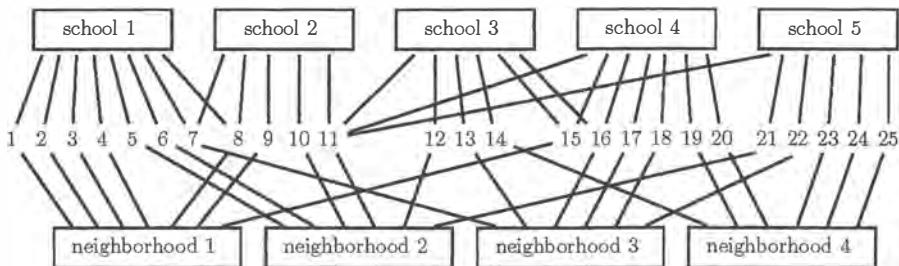


Figure 13.4: An example of pupils nested within multiple schools crossed with neighborhoods.

Table 13.4: A model with cross-classification and multiple membership for examination grades

Fixed effect	Model 7		Model 8	
	Coeff.	S.E.	Coeff.	S.E.
γ_0 Intercept	6.36	0.03	6.40	0.02
γ_1 Pretest			0.031	0.002
γ_2 SES			0.068	0.011
γ_3 Advice			0.053	0.010
γ_4 Ethnicity			-0.073	0.028
Random part	Var. comp.	S.E.	Var. comp.	S.E.
<i>Crossed random effect:</i>				
$\tau_w^2 = \text{var}(W_f)$ primary school	0.006	0.005	0.003	0.003
<i>Level-two random MM effect:</i>				
$\tau_0^2 = \text{var}(U_{0j})$ secondary school	0.065	0.014	0.033	0.007
<i>Level-one variance:</i>				
$\sigma^2 = \text{var}(R_{ij})$	0.395	0.010	0.331	0.008

the multiple membership model, where the weights go with the U_{0j} and, of course, not with the W_{0k} .

Example 13.3 Multiple membership in secondary schools cross-classified with primary schools.

We continue with the multiple membership model from the previous example, but now also simultaneously model the sustained effects of the primary schools. In Model 7 there are no predictor variables, and in Model 8 we once again include predictor variables. The results are presented in Table 13.4.

The table shows that the parameter estimates have not changed much as compared to the cross-classification model results presented in Table 13.1 and the multiple membership model results presented in Table 13.3.

The multiple membership multiple classification models can also be extended to situations with explanatory variables having random slopes (for further details see, for instance, Beretvas, 2011), and adding these effects may improve the last model as presented in Table 13.4. Models have also been developed for dependent variables with nonnormal distributions, as is discussed in the overview given by Rasbash and Browne (2008).

13.5 Glommary

Cross-classification or multiple classification. A situation where levels are not nested in each other. For example, lower-level units belong to higher-level units of two or more different but intersecting populations of interest. An example of such populations is given by pupils in neighborhoods and in schools, where pupils from the same neighborhood may attend different schools that also are attended by pupils from other neighborhoods.

Crossed random factor. A random factor that is crossed with another random factor.

Classification graph. A figure depicting the nesting of lower-level units in higher-level units.

Multiple membership. A situation where lower-level units belong to more than one higher-level unit of one population of interest. An example is given by pupils migrating from one secondary school to another.

Membership weight. A weight, for instance proportional to the length of membership duration, by which a lower-level unit is assumed to be affected by a higher-level unit.

Multiple membership multiple classification. A situation where lower-level units belong to more than one higher-level unit of one population of interest, and at the same time belong to higher-level units of one or more different but intersecting populations of interest.

14

Survey Weights

Some data collection methods use complex surveys, that is, surveys which cannot be regarded as simple random samples. In a *simple random sample*, which may be regarded as the standard statistical data collection design, each subset of the population having the same size has the same probability of being drawn as the sample.¹ In Chapter 2 we saw that two-stage samples and cluster samples are quite natural sampling designs for multilevel analysis. Another sampling method is stratification, where the population is divided into relatively homogeneous subsets called strata and samples are drawn independently within each stratum, with sampling proportions which may differ between strata. Many big surveys use a combination of stratification and multistage design. This will imply that elements of the population may have different probabilities of being included in the sample – these probabilities are known as the *sampling probabilities* or *inclusion probabilities*. This must be taken into account in the data analysis.

OVERVIEW OF THE CHAPTER

This chapter starts by discussing the choice between model-based and design-based approaches to inference. In brief, the model-based approach consists of incorporating aspects of the design (variables used for stratification, etc.) as far as necessary into the model, and proceeding with a hierarchical linear model in the usual way; design-based inference uses special techniques incorporating weights based on the sampling probabilities. We shall argue (Section 14.3) in favor of pursuing a model-based approach as much as possible, because this approach is more in line with the primary aims of analytic social science and because there are still issues with the procedures developed for multilevel design-based inference, in particular for small and medium sample sizes.

Section 14.1 introduces in general terms the concepts of model-based and design-based inference, and the debate – which has been more fervent among statisticians than among social scientists – about the choice between these modes of inference. In Section 14.2 the use of weights is discussed, and the reader is warned about the distinction between survey weights and precision weights. While these two first sections are about surveys in general, the following sections focus on multilevel analysis. Section 14.3 presents a variety

¹The distinction between sampling with and without replacement is not discussed here.

of methods for assessing whether a model-based approach is appropriate in a survey study with a multilevel data structure. Then in Section 14.5 the main methods proposed for multilevel model-based inference are discussed.

This chapter is about statistical methods that are still being developed and discussed, and therefore presents a less clear-cut account than some of the more basic chapters.

14.1 Model-based and design-based inference

14.1.1 Descriptive and analytic use of surveys

Surveys can have many purposes, and when choosing between methods of analysis it is helpful to distinguish between the descriptive and the analytic use of surveys (e.g., Pfeffermann, 1993). A survey is a study of a finite population, and *descriptive inference* is the estimation and testing of descriptive parameters of the population, as is done in official statistics. Consider, for example, the population of all pupils enrolled in a given type of school at a given date in some country. An example of such a descriptive parameter is the proportion of pupils who fail a specific test in mathematics. *Analytic inference*, on the other hand, is concerned with questions about the way in which a variable of interest depends on, or is associated with, other variables. Continuing the example, an analytic question would be how failing in mathematics depends on the gender of the pupil and of the teacher, not only in this particular finite population but more generally for comparable pupils in comparable schools. Here the researcher wishes to generalize to a larger population, which may be hypothetical and is usually rather vaguely defined – as signaled by the word ‘comparable’ in the example. The difficulty of describing the population to which one wishes to generalize is compounded in the multilevel case by the fact that there are several populations – in most cases, one for each level. But the difficulty in pinning down the population is not a grave concern because in practice we can live with a bit of vagueness.

When working with complex surveys, the probabilistic nature of statistical inference can be based on two approaches. One is the *model-based approach*, which is the basis of this book and of the majority of statistical texts. Here the researcher makes the assumption that the data can be regarded as the outcome of a probability model with some unknown parameters. Note that the word ‘model’ here refers not to the theoretical model from a social science viewpoint, but to what also might be called the ‘data generating mechanism’ – more precisely, the probabilistic mechanism that could have generated the values of the dependent variable, conditional on the explanatory variables and the multilevel structure. In linear regression-type models like the hierarchical linear model, the probabilistic element is represented by the residuals (sometimes called ‘errors’) which are the deviations between the model – which is always idealized – and the data.² The probability model is defined by the assumptions about the simultaneous probability distribution of all residuals occurring in the model. A discussion about the various types of probability models in statistical inference is presented by Cox (1990). Here we can take the view that randomness represents the influence of variables not included in the model together with the approximate nature of assumptions of linearity, measurement error, etc.

²For generalized linear models (e.g., multilevel logistic regression), it may be more enlightening not to talk about residuals at the lowest level, but about the conditional distribution of the data given the linear predictor.

In the case of the two-level hierarchical linear model, for instance, the probability model is expressed as in Chapter 5 by the assumptions that the level-one residuals are independent and identically distributed across level-one units, that the vectors of random intercepts and slopes are independent and identically distributed across level-two units, that these have normal distributions, that there is a linear dependence of the dependent variable on some array of explanatory variables, etc. These are assumptions that should *a priori* be plausible approximations when carefully considering the phenomena under study, and that can be tested (cf. Chapter 10), although uncertainty will always remain. A model never is more than an approximation.

The other is the *design-based approach*. Here the inference is based on the probability mechanism used for the sample selection. Probabilities that elements of the population are included in the sample may differ between elements; for example, in a survey of the adult population one might wish to oversample employed individuals, that is, give them a higher probability of inclusion than nonemployed persons. In the ideal case the design-based model is totally convincing because it only expresses the rules used in the data collection; in practice there will be some give and take because of nonresponse, imperfections in the sampling frame, and other so-called nonsampling errors. The inference is from the data to the finite population. In the boundary case where the data constitute a census, that is, the population is observed in its entirety, there is no need for statistical modeling because everything is perfectly known. The finite population is less idealized and less far away than the infinite population of the probability model, but regularities of human and social behavior belong to the realm of the idealized model rather than the finite population. For the finite population no assumptions about, for example, normal distributions need to be made; indeed no such assumptions *can* be made because for normal distributions the number of possible values is infinite. A crucial element in the design-based approach is the use of weights. In a one-level design, denote by π_i the probability that population element i is included in the sample. Then the data for this unit, if indeed included in the observed sample, will be weighted by a factor $w_i = 1/\pi_i$. This weighting counterbalances the lower frequency of observation of elements with a low inclusion probability.

It may be noted that in many introductory textbooks there is confusion because statistical modeling is often argued by underlining the importance of random sampling combined with making assumptions about independent and normally distributed residuals, thereby confounding the distinction between design-based and model-based inference.

This book is aimed at researchers who wish to obtain knowledge about social and behavioral processes and mechanisms, with some degree of generalizability beyond the data set under consideration. Thus it is oriented naturally toward model-based inference. However, a survey in which inclusion probabilities are not constant might give a biased representation of these social and behavioral processes, and the question then is how to analyze such a data set to obtain unbiased, or approximately unbiased, conclusions. This issue has led to a debate between statisticians favoring a design-based and those favoring a model-based approach. Although the sharpness of the debate has attenuated, there are still clear differences of opinion. An overview written for social scientists, with an interesting presentation of the history of this debate going back to Fisher and Neyman, is given by Sterba (2009). Those who would like to consult some of the original sources could have a look at DuMouchel and Duncan (1983), Kish (1992), Pfeffermann (1993), Smith (1994), Little (2004), Gelman (2007), and articles cited therein.

The contrasting positions in their barest forms are as follows. Proponents of a model-based approach (Bayesians³ as well as frequentists) say that if the model is true and the sampling mechanism is independent of the residuals in the probability model, then the sampling design is irrelevant and taking account of the sampling design entails a loss of efficiency (Kish, 1992). This efficiency loss will be serious if the weights w_i are highly variable across population elements: a few sample elements will then be weighted very strongly. This also implies that the residuals associated with these elements will be weighted strongly, and if residual variance is constant this will lead to large standard errors. Proponents of the design-based approach, on the other hand, say that one never can be confident that the model is true, and therefore a model-based estimator might be seriously biased and thus worthless. Bias here means a misrepresentation of the true data-generating mechanism, induced by the distortions due to the survey design. Design-based estimators yield protection against such distortions, and are possible because the design, and therefore the potential distortions, are exactly known as a function of the design variables (e.g., DuMouchel and Duncan, 1983).

Some degree of reconciliation between the opposing positions has taken place (e.g., Pfeffermann, 1993; Smith, 1994). Modelers realize that it may be important to include design-related elements in the models. The use of the hierarchical linear model for two-stage samples is an example (see, for example, Little, 2004). Survey-oriented statisticians have developed methods incorporating models. In this chapter we follow such a compromise approach, but with an argued preference in favor of model-based inference, as will be made clear below.

14.2 Two kinds of weights

Weights in statistical analysis can be of several kinds, and the distinction between these is not always clear in statistical software manuals. Avoiding confusion between the two main kind of weights is essential, and this is the purpose of this section. We start by treating this distinction for single-level linear regression analysis. More extensive treatments are given by DuMouchel and Duncan (1983) and Kish (1992).

Sampling weights or *survey weights* are used in surveys from finite populations and reflect inversely the probability that any particular population element is included in the sample. These are constant for simple random samples, and then they may safely be ignored; for complex samples, such as stratified and two-stage samples, they usually are nonconstant and the issue of how to take this into account is the topic of this chapter. For example, a sample of adults could be stratified with respect to employment status, with nonemployed people having a higher inclusion probability than employed.

In a one-level design where π_i is the probability that population element i is included in the sample, the (nonstandardized) weight is $w_i = 1/\pi_i$. Cases that are included with a low probability are in a part of the population that is under-represented in the sample, and must receive a larger weight to obtain unbiased estimators of properties of the population. Large weights will be associated with large uncertainty, because one sampled individual is used to represent a large portion of the population. Denote the parameters by β and the fitted

³See Section 12.1.

value for case i by $\hat{y}_i(\beta)$. The *probability weighted estimate* (PWE) for β is then obtained by finding the value of β for which

$$\sum_{i \in \text{sample}} w_i (y_i - \hat{y}_i(\beta))^2 \quad (14.1)$$

is as small as possible.

For the case where only a population mean is being estimated, that is, β is one-dimensional and $y_i(\beta) = \beta$, this reduces to

$$\hat{\beta} = \frac{\sum_i w_i y_i}{\sum_i w_i} \quad (14.2)$$

with estimated standard error

$$\text{S.E.}^p(\hat{\beta}) = \sqrt{\frac{\sum_i w_i^2 (y_i - \hat{\beta})^2}{\left(\sum_i w_i\right)^2}}. \quad (14.3)$$

This design-based standard error, that is, estimator for the sampling standard deviation of $\hat{\beta}$ under random sampling according to the design with inclusion probabilities π_i , is a special case of the formula presented, for example, in Fuller (2009, p. 354).

Precision weights, also called analytical weights, indicate that the residual standard deviations of some cases are larger than for others, and estimates will be more precise if cases with higher residual standard deviation get lower weight. Denote in a one-level design the residual standard deviation for case i by s_i . One reason why residual standard deviations might be nonconstant is that a case in the data set might be an average of a variable number of more basic units; then the weights may be called *frequency weights*. Another possible reason is that the dependent variable might represent a count or quantity, and higher quantities mostly are naturally associated with larger variability. The (nonstandardized) weight associated with standard deviation s_i is $w_i = 1/s_i^2$. Here, larger weights reflect smaller standard errors (i.e., less uncertainty); this is opposite to the case explained above of sampling weights. Here also, the weighted parameter estimate is obtained by finding the minimum of

$$\sum_{i \in \text{sample}} w_i (y_i - \hat{y}_i(\beta))^2 \quad (14.4)$$

as a function of β . This is called the weighted least squares (WLS) estimator.

The PWE estimator (14.1) and the WLS estimator (14.4) are defined by the same expression, but they are justified differently. Since they are based on different assumptions, the expressions usually given for their variances are different. The appendix to this chapter (Section 14.6) contains the matrix expressions for these variances. For the case of estimating a sample mean, if the variance for Y_i is $\text{var}(Y_i) = \sigma^2/w_i$, then the WLS estimator (14.4) again reduces to (14.2). The standard error can then also be estimated by (14.3), which here is a simple (one-dimensional) case of the sandwich estimator proposed by Liang and Zeger (1986); see Section 12.2.

This estimator (14.3) for the standard error is heteroscedasticity-consistent, that is, for large samples it will give good estimates even if the variables Y_i have nonconstant variances. However, the estimated standard error often given is

$$S.E.^W(\hat{\beta}) = \sqrt{\frac{\sum_i w_i (y_i - \hat{\beta})^2}{\sum_i w_i}}, \quad (14.5)$$

This is correct and efficient if indeed $\text{var}(Y_i) = \sigma^2/w_i$, but not heteroscedasticity-consistent: it is not a reliable standard error if the variances are not inversely proportional to the weights.

Precision

The reason for using the WLS estimator (14.4) with precision weights is to obtain high precision.⁴ In the case of survey modeling, however, the PWE estimator (14.1) is used to avoid bias, and its precision is not necessarily optimal. Under the usual assumption of *homoscedasticity* (constant residual standard deviations), there might be some cases with large sampling weights w_i which happen to have large residuals (which means that through random causes their fit with the model is less good than for most other cases), and the minimization of (14.1) will give such cases a large influence. The extent of inefficiency of the PWE estimator (14.1) in the homoscedastic case can be expressed by the *effective sample size* (Pothoff et al., 1992), defined as

$$n_{\text{eff}} = \frac{(\sum_i w_i)^2}{\sum_i (w_i^2)}, \quad (14.6)$$

It is called this because when estimating by the PWE estimator (14.2) the mean of some variable which has variance σ^2 independently of the sampling weights w_i , the variance of the estimator is σ^2/n_{eff} , just as if it were the regular mean of a sample of size n_{eff} . The effective sample size is strictly smaller than the regular sample size, unless the weights are constant; therefore in the homoscedastic situation the use of sampling weights leads to a loss of precision, and this can be serious if the variability of the weights is large, as will be expressed by a small value for the effective sample size.⁵

If the residual variance is nonconstant (*heteroscedasticity*), however, the effect on the estimation variance of using the weighted estimation (14.1) will be different. The assumptions underlying survey weights and those underlying precision weights are independent, and therefore they can both be true – in which case the use of sampling weights yields a gain, rather than a loss, of efficiency compared to unweighted estimation.

Bias

The reason for using sampling weights is to avoid bias. A sampling design with heterogeneous inclusion probabilities will yield a sample that is a biased representation of the

⁴In other words, low mean squared error.

⁵It should be noted that this measure for the design effect and its interpretation of the loss of precision applies only to the case of estimating a mean of a homoscedastic population; for the estimation of regression coefficients the formulas are more complicated.

population. This does not, however, necessarily lead to bias in the estimators for the parameters. The hierarchical linear model is a model for the dependent variable, conditional on the explanatory variables and the multilevel structure. If the predictor variables are represented differently in the sample than their distribution in the population – which is one kind of bias – it does not follow that there must be bias in the estimators of the parameters of the hierarchical linear model. The parameters are biased only if the distribution of the residuals is affected by the sampling design. Then the survey design is said to be *informative*. If the residuals in the model are independent of the sampling design and of the sampling weights, that is, the model is correctly specified given all variables including the design variables, then the use of weights is superfluous. In multilevel analysis of survey data, if we can be confident of working with a well-specified hierarchical linear model and the sample design is unrelated to the residuals, it is better to take no explicit account of the survey design when doing the analysis and proceed as usual with estimating the hierarchical linear model. The difficulty is, of course, that we never can really be sure that the model is well specified.

As an example of the meaning of this type of bias, consider a multilevel study of pupils in schools with language achievement as the dependent variable, and socio-economic status (SES) as one of the explanatory variables. Assume that the sampling design is such that the inclusion probability of schools is likely to be related to the ethnic composition of the schools, but no data about ethnic composition are available. Assume also that language achievement depends on ethnic background of the pupil, and SES is the main explanatory variable included that is related to ethnic background. The hierarchical linear model for language achievement is then misspecified because ethnic background is not included, and this will mainly affect the estimate for the effect of SES. The SES variable in this data set and model reflects a variable where different ethnic subgroups of any given value on SES are distributed in proportions that differ from the true population proportions, thus leading to a biased parameter estimate for SES. This argument hinges on the fact that there is an omitted variable, correlated with an included predictor variable, and correlated with the inclusion in the sample.

Clearly, the most desirable solution here would be to observe the ethnic background of pupils and include it in the model. A second solution is to come up with a variable that reflects the sampling design and that is closely related to the ethnic composition of the school, and control for this variable. A third solution is to weight schools according to their inverse inclusion probability. In the rest of this chapter we shall discuss such solutions in more detail.

14.3 Choosing between model-based and design-based analysis

This section presents arguments and diagnostics for choosing between a model-based and a design-based analysis of a given multilevel data set. It is assumed that the purpose of the statistical inference is analytic: we wish to find out how a dependent variable depends on relevant explanatory variables. This is expressed in a hierarchical linear model for the dependent variable, including the relevant explanatory variables, which will be called the *model of interest*. This may be either a model given on the basis of theory, or a model where some of the details are still to be specified depending on the data. The design is assumed

to depend on a number of variables called the *design variables*. These are the variables on which the inclusion probabilities depend. With a *design-based approach to analytic inference* we refer to an approach that analyzes the model of interest while taking the design into account. Thus, given that it is about analytic inference, this approach is based on a combination of model-based and design-based considerations. For single-level studies we hinted at such a combined approach by presenting the probability weighted estimator (14.1) based on fitted values, which presupposes a model; for multilevel studies this will be discussed further in Section 14.5.

Our position is that it will be preferable, if possible, to carry out a model-based analysis, possibly according to a model that is augmented by also using some of the design variables; and that a design-based analysis will be done only if there are clear arguments against a model-based analysis. The arguments for this preference are as follows:

1. Finding a well-fitting model for the dependent variable is the main purpose of analytic inference. In the model-based approach, augmented by design variables, one tries to find those elements of the design variables that may contribute to understanding the dependent variable. This serves the main purpose more directly than a design-based approach where one tries to ‘weight the design away’.
2. Most researchers reading this book will be better acquainted with the model-based than with the design-based approach, and they will be able to carry out a model-based analysis with more confidence and less risk of error. In this sense, the model-based analysis has the virtue of simplicity.
3. The methods developed for the design-based analysis of multilevel data structures are good for estimation of fixed effect parameters but less so for the estimation of variance parameters and standard errors and for hypothesis tests unless sample sizes are fairly large.

Thus we shall lay out an approach where the design variables are scrutinized for their effect on the dependent variable, and it is investigated whether it is sufficient to analyze the model of interest, possibly extended with design variables, by a regular hierarchical linear model; and where an approach that analyzes the data in a way that also takes account of the design, by using survey weights, is followed only if there are reasons not to trust a model-based analysis. We could say that we wish to assess whether the sampling design is *noninformative* (Pfeffermann, 1993), meaning that the residuals in the statistical model are independent of the variables used for the survey design. In that case we can proceed with the analysis without taking the design into account.

It should be noted that the multilevel analysis according to the hierarchical linear model employs a model which already accounts for one element present in many sample designs, namely, the clustering present in cluster samples and multi-stage samples. But different sampling probabilities occurring, e.g., in stratified or post-stratified sampling are not accounted for in a straightforward way, and how to do this is the issue of this chapter.

14.3.1 Inclusion probabilities and two-level weights

The treatment focuses on two-level models. We shall sometimes use the term *clusters* to refer to level-two units, and use *cluster size* for the sample size n_j for a given level-two unit j .

Given the two-level structure, we assume a two-stage sample design, where clusters are chosen independently with some probability, and given that a cluster j has been selected, a sample of level-one units within this cluster is chosen. The *inclusion probabilities* are defined as

$$\pi_j = \text{inclusion probability for level-two unit } j; \quad (14.7a)$$

$$\pi_{ij} = \text{inclusion probability for level-one unit } i, \quad (14.7b)$$

given the inclusion of cluster j .

It is possible that $\pi_j = 1$ for all j : then all clusters in the population are included in the sample. An example is a survey in countries as level-two units, where all countries satisfying some criterion (defining the population) are included in the study, and within each country a sample of the population is taken. A different possibility is $\pi_{ij} = 1$ for all i and j ; then the sample design is called a cluster sample, and clusters are observed either completely or not at all. An example is a survey of schools as level-two units, where within each sampled school all pupils are observed. The marginal probability of observing level-one unit i in cluster j is given by the product,

$$\pi_j \pi_{ij} = \text{inclusion probability of level-one unit } i \text{ in cluster } j.$$

The *weights* or *design weights* are the inverse of the inclusion probabilities. Thus, the weights are defined by

$$w_j = \frac{1}{\pi_j}, \text{ weight for level-two unit } j; \quad (14.8a)$$

$$w_{ij} = \frac{1}{\pi_{ij}}, \text{ weight for level-one unit } i \text{ in level-two unit } j. \quad (14.8b)$$

To use weights in two-level models, the separate sets of weights at level one and level two are needed, corresponding to the separate inclusion probabilities at level one and level two.

From these weights the *effective sample sizes* can be calculated:

$$N^{\text{eff}} = \frac{(\sum_j w_j)^2}{\sum_j (w_j^2)}, \text{ effective sample size at level two}; \quad (14.9a)$$

$$n_j^{\text{eff}} = \frac{(\sum_i w_{ij})^2}{\sum_i (w_{ij}^2)}, \text{ effective sample size at level one for cluster } j. \quad (14.9b)$$

The effective sample size is defined such that a weighted sample gives the same amount of information as a simple random sample with sample size equal to the effective sample size.

The ratios of effective sample sizes to actual sample sizes are called the *design effects* at level two and level one,

$$\text{deff}_2 = N^{\text{eff}}/N \quad \text{and} \quad \text{deff}_{1j} = n_j^{\text{eff}}/n_j. \quad (14.10)$$

The design effects give a first indication of the potential loss of statistical efficiency incurred by following a design-based rather than model-based analysis (but see footnote 5 on p. 221).

In single-level studies the scales of these weights are irrelevant,⁶ that is, all weights could be multiplied by the same number without in any way changing the results of the analysis. In multilevel designs scaling the level-two weights still is irrelevant, but the scale of the level-one weights is important, and the literature is still not clear about the best way of scaling.

Pfeffermann et al. (1998) propose two methods of scaling. These are used in design-based estimation for multilevel designs (see below in Section 14.5). The first is

$$\text{method 1: } w_{ij}^* = s_j^{(1)} w_{ij} \quad \text{with } s_j^{(1)} = \frac{n_j^{\text{eff}}}{\sum_i w_{ij}}, \quad (14.11a)$$

which yields scaled weights summing to the effective sample size n_j^{eff} ; this is called ‘method B’ by Asparouhov (2006). The second is

$$\text{method 2: } w_{ij}^o = s_j^{(2)} w_{ij} \quad \text{with } s_j^{(2)} = \frac{n_j}{\sum_i w_{ij}}, \quad (14.11b)$$

yielding weights summing to the actual sample size n_j , and called ‘method A’ by Asparouhov.

14.3.2 Exploring the informativeness of the sampling design

If there is an association between the sampling design and the residuals in the hierarchical linear model, then the sampling design is informative (Pfeffermann et al., 1996) and a model-based approach risks being biased and inconsistent, that is, the parameter estimates may be a misrepresentation of the true parameters in the hierarchical linear model even in large samples. Here we propose a variety of methods that can be followed to assess the association of the sampling design with the model of interest. They do not have the primary aim of obtaining a yes/no verdict about the informativeness of the design for the given model of interest. Their purpose is rather to explore how the design variables may be associated with the model of interest, to get ideas about how to extend the model of interest with relevant variables reflecting the design, and, after this exploratory phase in which perhaps the model of interest was extended, to assist the researcher in deciding between a model-based and a design-based analysis. In any single study it would probably be overkill to use all methods presented here. Which of these methods are useful will depend on the data structure and the aims of the analysis, to be decided by the insight and experience of the researcher.

1. *Look at the variability of the weights.* Explore the variability of the weights, for the level-two weights w_j as well as for the level-one weights w_{ij} within each sampled cluster j . This is done in the first place by calculating effective sample sizes at level two, and at level one for each cluster, defined above in (14.9). If actual cluster sizes n_j are strongly variable, it may be illuminating to plot the design effects (14.10) as a function of actual cluster sizes.

Depending on the situation, more details of the weight distribution may be helpful, for example, by making boxplots. High variation of weights and low design effects

⁶It is assumed here that finite population corrections are ignored.

are warning signals: they imply that the design is far from a random sample, which leads to the possibility of considerable biases of model-based estimators, and to low efficiency of design-based estimators.

2. *Consider adding design variables to model of interest.* In the ideal case, all variables determining inclusion probabilities – the *design variables* – are known and are part of the data set. This allows a much clearer and more satisfactory analysis than knowing only the weights. However, especially for large surveys, it is not uncommon that information about the design is only partial, for example, for confidentiality reasons or when weights have been determined not only because of nonconstant sampling probabilities but also to correct for nonresponse.

If the design variables are available to the researcher, then they may be included in the model; it can be tested whether they affect the dependent variable and whether the other results are sensitive to their inclusion. This makes sense especially as additions to a model of interest that is already rich enough in explanatory and control variables; the only concern in this exploration is whether the design variables have effects over and above the variables in the model of interest.

The design variables should be included not only as main fixed effects but also in interactions with other variables and possibly in the random part of the model. It is advisable to test their interactions with the main explanatory variables in the model, and also to test whether they have random slopes or are associated with heteroscedasticity (cf. Chapter 8). General rules cannot be given here because everything depends on the specific properties of the design, the data structure, and the model of interest. For example, if the design consists of a stratification (i.e., distinct subsamples) at level one or level two, the best approach may be to estimate the hierarchical linear model separately for each of the subsamples and assess the importance of the differences between the results of the subsamples. Then in a next step this can lead, if desired, to a common analysis where a number of main effects of the subsamples and of interactions are included to represent these differences. Another possibility here is to include the subsamples (strata) as an additional level in the model and, given these, random main effects and random slopes according to what is a good model specification. This use of random effect models is also discussed by Little (2004, equation (14)), and Gelman (2007).

If the researcher is certain that the design variables have been included in a satisfactory way, then a model-based analysis can be pursued further. In practice, of course, one never is really sure about this, especially since complicated interactions and nonlinear effects might possibly be involved. To explore whether the model differs in ways associated to the design weights, the following procedures can give some insights.

3. *Apply the hierarchical linear model to parts of the data set differing with respect to inclusion probabilities or other design factors.* First we present one way to investigate the influence of the design on the estimation of the model of interest which is easily carried out and can give insight in the magnitude of this influence.

The procedure is to split the data set into parts according to the weights at level one and those at level two, analyze each part separately by the hierarchical linear model according to the model of interest, and assess the magnitude of the differences in

results between the various parts. How to make this split depends on the sample sizes and the distribution of the weights; a possibility is the following. Divide the level-two units (clusters) into two or three groups based on the level-two weights w_j ; within each cluster, divide the level-one units into two or three groups based on the scaled level-one weights w_{ij}^* . (The choice between scaling methods 1 (14.11a) and 2 (14.11b) is rather arbitrary here and in most cases will not make much difference; we suggest method 1.) The combination of these two divisions splits the data set into four to nine (or a different number, depending on how the split is done) disjoint parts, and for each part the model of interest is analyzed using a regular multilevel model. Each of the resulting sets of parameter estimates reflects a part of the populations at levels one and two that was sampled with different inclusion probabilities. If these sets of results are more or less the same (given their standard errors and given what is deemed to be an important difference), then it is unlikely that the design has a large impact on the estimation of the model of interest. On the other hand, if the sets of results do differ from each other, then there is evidence for such an impact. The type of difference may give some insight in the type of impact, for example, whether it is a consequence of mainly level-two weights, mainly level-one weights, or both. In some cases this may even lead to finding design-related variables that can be added to the model of interest, after which there might be no further appreciable impact of the design on the estimation results.

4. *Add weight variables to the model.* Noninformativeness of the design means that the design variables are independent of the residuals. This can be tested by using the weights as predictor variables added to the hierarchical linear model (cf. DuMouchel and Duncan, 1983). Note that we have two sets of weights, at level one and level two. Moreover, the level-one weights can be scaled and there are no convincing guidelines on how to scale. We tentatively propose to use scaling method 1.

The procedure is as follows. For the level-two weights w_j and the scaled level-one weights w_{ij}^* of formula (14.11a), we test the main effects as well as all interactions with explanatory variables in the model of interest. This test is carried out by adding the fixed effects of these variables to the hierarchical linear model. Since this leads to a multiparameter test, a straightforward way of doing this is by using the ML estimators and the deviance test (Section 6.2). If this test is not significant, there is no evidence for bias of the model-based estimator.

This procedure can be extended by testing heteroscedasticity of the level-one residuals as a function of the level-one and level-two weights according to the methods of Section 8.1. It follows from Section 14.2 that if residual variance increases as a function of the weights, then the design-based estimator is particularly inefficient; if it decreases as a function of the weights, then the design-based estimator may be relatively efficient.

Depending on the research questions and variables at hand, these tests could also be applied to nonlinear transformations of the weights. If a good search for effects of weight-related effects on the dependent variables does not lead to findings of such effects, it may be justified to continue with a model-based analysis that does not take the design into account.

Depending on the design, it can be helpful to compare estimates that are purely model-based with estimates that also take design-based estimates for meaningful parts of the data set, or for the entire data.

5. *Compare model-based and design-based estimators separately for each cluster.* If a great deal of detail is required with respect to the consequences of the sampling design within the clusters, and if the clusters are reasonably large (e.g., a difference of 30 or more between each cluster size and the number of level-one predictors), the comparison between the model-based and the design-based analysis can be done on a cluster-by-cluster basis. If the model of interest is the regular hierarchical linear model (5.15) without level-one heteroscedasticity,

$$Y_{ij} = \gamma_0 + \sum_{h=1}^r \gamma_h x_{hij} + U_{0j} + \sum_{h=1}^p U_{hj} x_{hij} + R_{ij},$$

then the model-based estimator within each cluster is the ordinary least squares (OLS) estimator. After all, in line with (5.1), the within-cluster model is just a single-level linear regression model. The within-cluster OLS estimator estimates the regression parameter including the random slope (if any), that is, the parameter $\gamma_h + U_{hj}$ for the variables $h = 0, 1, \dots, p$ which do have a random slope, and γ_h for variables $h = p+1, \dots, r$ which do not. Denote this estimator (a vector with $r+1$ components) by $\hat{\beta}_j^{\text{OLS}}$ and its covariance matrix (which has the squared standard errors on the diagonal) by Σ_j^{OLS} . Denote the design-based estimator (14.1) by $\hat{\beta}_j^W$ with covariance matrix Σ_j^W . (For this single-level situation referring to a single cluster, the scaling of the weights w_{ilj} is immaterial.)

If the model is valid and the design variables are uncorrelated with the level-one residuals R_{ij} , then the difference $\hat{\beta}_j^{\text{OLS}} - \hat{\beta}_j^W$ is just random noise; on the other hand, if the design variables are correlated with the level-one residuals, then this difference estimates the bias of the OLS estimator. A theoretically interesting property is that, if this model is valid, the covariance matrix of the difference is the difference of the covariance matrices:⁷

$$\text{cov}(\hat{\beta}_j^W - \hat{\beta}_j^{\text{OLS}}) = \Sigma_j^W - \Sigma_j^{\text{OLS}}. \quad (14.12)$$

DuMouchel and Duncan (1983) discussed this and showed that the test for the difference $\hat{\beta}_j^W - \hat{\beta}_j^{\text{OLS}}$ can also be carried out as follows in a linear OLS model. Compute the product of all level-one variables x_{hij} with the weights w_{ilj} , and test in an OLS model the effect of the weight w_{ilj} itself and all r products $x_{hij}w_{ilj}$, controlling for X_1, \dots, X_r . This yields an F -test with $r+1$ and $n_j - 2(r+1)$ degrees of freedom which is exactly the same as the F -test for the difference between the design-based and the model-based estimators. It can be regarded as the test for the main effect of the weight variable, together with all its interactions with the explanatory variables. A version of this test for the case of multilevel models for discrete outcome variables was developed by Nordberg (1989).

⁷This relation does not hold in general. It is valid here because of the independence between the OLS estimator and the difference $\hat{\beta}_j^W - \hat{\beta}_j^{\text{OLS}}$. The matrix expression for (14.12) is $\sigma^2((X'WX)^{-1}(X'W^2X)(X'WX)^{-1} - (X'X)^{-1})$; cf. (14.20).

This yields a test outcome for each of the N clusters. This information can be used for diagnosing the noninformativeness of the sampling design in all clusters individually. If one also wishes to have an overall assessment, the following procedure can be used to combine the N tests into one. It consists of two steps. Suppose that the test for cluster j yields the F -statistic F_j , and denote the degrees of freedom by df_1 for the numerator and df_{2j} for the denominator. In the first step, the F_j values are transformed to an approximate standard normal value by Fisher's Z transformation⁸ (see Johnson et al., 1995, Section 27.5) defined by

$$Z_j = \frac{\ln(F_j) + df_1^{-1} - df_{2j}^{-1}}{\sqrt{2(df_1^{-1} + df_{2j}^{-1})}}, \quad (14.13)$$

where $\ln(F_j)$ is the natural logarithm of F_j . In the second step, add the normalized values

$$Z = \frac{\sum_j Z_j}{\sqrt{N}}; \quad (14.14)$$

this test statistic can be tested in a standard normal distribution. Another procedure for combining the tests is Fisher's combination of p -values (3.35).

These F -tests are exact (if the level-one residuals are normally distributed with constant variance) and therefore also are valid for small sample sizes n_j . (Fisher's combination of p -values also is exact, but Fisher's Z transformation is an approximation.)

If the denominator degrees of freedom are reasonably large (not less than 20, preferably greater than 30), the differences between the OLS and the weighted estimates of the regression coefficients are reliable enough to be used for exploration of the effect of the within-group designs. For the most important level-one variables, these differences can be plotted as a function of the OLS estimate or other relevant cluster characteristics, and this can give insight in the pattern of how the level-one design is associated with the model of interest.

If the combination of the F -tests is not significant, this is an indication that in the analysis of the model of interest it is not necessary to take the level-one sampling design into account, so that only the level-two sampling design remains. The case of having only level-two weights and no (that is to say, constant) level-one weights is considerably simpler than the case of variable weights at both levels, so that this will give an important simplification.

It should be realized, however, that these F -tests only look at the estimated values and not, for example, at the variances and standard errors, which might also be affected by the design; and that a significant test result can be interpreted as a signal of informativeness, but nonsignificance cannot be interpreted as a reliable confirmation of noninformativeness. Therefore, these procedures can be very helpful in exploring

⁸There are two different Fisher Z transformations; the better-known of these is for the correlation coefficient, while this one is for the F distribution. See Fisher (1924) and Fisher and Yates (1963, p. 2) ('The quantity $z\dots$ '). For this reference we are indebted to sir David Cox.

the consequences of the design, but they cannot give clear confirmation that nothing is being overlooked.

6. *Test differences between model-based and design-based estimators.* The model-based and design-based estimators can both be calculated and compared. The test of the difference between these two estimators, explained above for the case of a single-level homoscedastic (OLS) model, can be generalized to testing this difference for any model and any design. With the currently available methods this does not lead to a reliable procedure in the multilevel case, however. Therefore, although it is meaningful to compare the two estimates, a formal statistical test is not available in practice. We explain the basic procedure and the associated difficulties, and then give a proposal which is practical but not a formal test.

Pfeffermann (1993) noted that the test of DuMouchel and Duncan (1983) is a special case of the Hausman (1978) principle of comparing an efficient estimator with a nonefficient one, and that (14.12) holds quite generally for large sample sizes. To explain this test, denote by $\hat{\gamma}^{\text{HLM}}$ the ML or REML estimator for the vector of fixed parameters γ in model (5.15) and by $\hat{\gamma}^W$ a design-based estimator. Similarly, denote the estimated covariance matrices of these estimators (with squared standard errors on the diagonal) by $\hat{\Sigma}^{\text{HLM}}$ and $\hat{\Sigma}^W$. The test statistic for comparing the two estimators is then

$$(\hat{\gamma}^{\text{HLM}} - \hat{\gamma}^W)' (\hat{\Sigma}^W - \hat{\Sigma}^{\text{HLM}})^{-1} (\hat{\gamma}^{\text{HLM}} - \hat{\gamma}^W) \quad (14.15)$$

and under the null hypothesis that the sampling design is noninformative, it has an asymptotic chi-squared distribution with $r + 1$ degrees of freedom, where $r + 1$ is the number of elements of parameter γ . This test can also be applied componentwise: if γ_h is the fixed coefficient of variable X_h (which is the intercept if $h = 0$),⁹ then

$$\frac{\hat{\gamma}_h^{\text{HLM}} - \hat{\gamma}_h^W}{\sqrt{\hat{\Sigma}_{hh}^W - \hat{\Sigma}_{hh}^{\text{HLM}}}} \quad (14.16)$$

can be used for testing the noninformativeness of the design specifically for coefficient γ_h ; under the null hypothesis of noninformativeness this has an asymptotic standard normal distribution.

A difficulty with these tests, however, is that the estimated covariance matrix of the weighted estimator may be quite unstable, especially for small or moderately large level-two units. For example, it is possible that the difference under the square root in the denominator of (14.16) is negative. This makes the tests unreliable. This difficulty was stressed by Asparouhov (2006). He proposed instead to use these calculations to calculate a measure for the informativeness of the sampling design, defined by

$$I_2 = \frac{\hat{\gamma}_h^{\text{HLM}} - \hat{\gamma}_h^W}{\sqrt{\hat{\Sigma}_{hh}^{\text{HLM}}}} \quad (14.17)$$

⁹Note the difference in notation: what in the clusterwise method was denoted by γ_j is the $(r + 1)$ -dimensional parameter vector for the j th cluster, whereas what is here denoted by γ_h is the single number which is the parameter for variable X_h .

This can be regarded as an effect size for the deviation between the model-based and the design-based estimators: when this difference is small, the difference between the two estimates is not important, whether or not it is significant.

We should note that the caveats, expressed at the end of item 4, against relying too strongly on testing differences between model-based and design-based estimates, apply also here.

14.4 Example: Metacognitive strategies as measured in the PISA study, USA 2009

As an elaborate example, this section is concerned with an analytic investigation of metacognitive strategies as measured in the international PISA study, using the data collected in 2009 in the USA (OECD, 2010). The dependent variable is indicated by the name METASUM in the PISA documentation. This is a measure of a metacognitive aspect of learning, namely the awareness of appropriate strategies to summarize information.

14.4.1 Sampling design

The survey design for the PISA 2009 study in the USA is a stratified two-stage sample. The first step was stratification by the combined classification of school type (private and public) and region (Midwest, Northeast, South, and West). Then within each stratum a two-stage sample was taken, in total of 5,233 students within 165 schools.

The numbers of sampled schools and students are listed in Table 14.1. It can be concluded that the number of private schools sampled is so small that meaningful inferences for the population of private schools, as distinct from the population of public schools, are hardly possible. It is possible to estimate a model for the public and private schools jointly, and test for differences between these two populations; but this test will have very low power in view of the small sample size for private schools.

Table 14.1: Numbers of sampled schools (left) and sampled pupils (right) per stratum, for PISA data, USA, 2009.

	Midwest	Northeast		South		West	
Public schools	38	1,264		26	732	55	1,776
Private schools	2	71		2	46	4	108

The data set contains two weight variables, W_FSTUWT at the student level and W_FSCHWT at the school level. The documentation (OECD, 2010, p. 143) mentions that these are to be transformed as follows. The student-level weights are to be divided by their school average:

$$w_{1ij} = \frac{n_j (W_FSTUWT)_{ij}}{\sum_k (W_FSTUWT)_{ik}} . \quad (14.18)$$

In other words, the sum $\sum_i w_{1ij}$ of the student-level weights per school is equal to the sample size n_j of this school. This is scaling method 2 of (14.11b). The school-level weights are to be normalized so that their sum over all students in the data set is equal to the total number of sampled students:

$$w_{2j} = \frac{M(W_FSCHWT)_j}{\sum_h n_h (W_FSCHWT)_h} \quad (14.19)$$

The technical documentation for the PISA 2009 survey was unavailable at the time of writing (early 2011), and therefore the technical report of the PISA 2006 survey (OECD 2009) is used for further background information. This report indicates (Chapter 8) that within the strata, schools were sampled proportional to total enrollment of the schools, and within schools pupils were sampled randomly (with equal probability). The weights given in the data set reflect not only the selection probabilities but also adjustment for nonresponse (so-called post-stratification).

To investigate how the sampling design could or should be included in the analysis, we shall potentially consider the eight strata separately. This is because they were sampled independently; the public–private dimension is substantively meaningful and the geographic dimension is potentially meaningful; and the number (eight) of separate analyses is still just manageable.

As a first step we wish to gain some insight into the variability of the weights. We calculate the weights (14.18) and (14.19) and compute the design effects (14.10). It turns out that all level-one design effects are between 0.95 and 1. This means they are hardly variable, and unequal weights between pupils cannot lead to important adjustments or biases. The level-two design effects must be considered per stratum. They are listed only for the public schools in Table 14.2; these numbers would be meaningless for the private schools, given the low number of such schools. This means, for example, that, for the purpose of estimating the mean of a homoscedastic variable, the sample for public schools in the South and West has the precision of a sample which has less than 20% of the size of a simple random sample. The precise definition of homoscedasticity in this case is that the absolute deviation from the mean is unrelated to the sampling weights.

Table 14.2: Level-two design effects for PISA data (USA 2009).

	Midwest	Northeast	South	West
Public schools	0.49	0.74	0.16	0.17

Illustration: descriptive analysis

To illustrate the implications of the nonconstant sampling weights for schools, consider first the descriptive analysis of three parameters, for the population of public schools in the South: the proportion of schools in villages, in towns, and in cities. Table 14.2 reports the design effect for schools within this stratum as 0.16, which indicates rather high although not outrageous variability of sampling weights. Since sampling of schools was done with

probabilities proportional to school enrollment, and larger populations will be served by bigger schools, the school weights will be rather strongly associated with schools being situated in a village (positively) or in a city (negatively), and not or only weakly with being situated in a town.

We compare the design-based and model-based approaches. The design-based approach uses the probability weighted estimates (14.2) with the corresponding standard errors (14.3). The model-based approach makes the usual model assumption of an ‘independent and identically distributed random sample’, implying that the sample elements have a constant variance (homoscedasticity), for which the estimate of the population average is the unweighted sample mean with the usual standard error of the mean. Table 14.3 presents the results. The estimated proportions of schools situated in villages, towns, and cities are, respectively, 0.18, 0.65, and 0.17. Large schools are oversampled, leading to smaller average weights in cities, whereas small schools are undersampled, leading to larger average weights in villages. This is reflected by the sample proportions: relatively more sampled schools are in cities, relatively fewer in villages. The sample proportions for schools in villages and towns would be biased if considered as estimates for the population proportions.

Table 14.3: Design-based and model-based estimates for three proportions, for public schools in the South.

Parameter		Probability-weighted	Unweighted	
Proportion of schools situated in a	$\hat{\beta}^P$	S.E. ^P	\bar{y}	S.E.(\bar{y})
Village	0.18	0.09	0.11	0.04
Town	0.65	0.13	0.56	0.07
City	0.17	0.07	0.33	0.06

The estimated proportion in towns, 0.65, does not differ much from the sample proportion, 0.56. The difference is less than one standard error (0.13). This is in line with the small association between sampling weights and being situated in a town. The total sample size here (number of sampled public schools in the South) is 55 (see Table 14.1). If in a simple random sample of size 55 a proportion of 0.65 were observed, the standard error of the estimated proportion would be 0.06, against the standard error of the weighted estimator S.E.^P($\hat{\beta}^P$) = 0.13. This indicates the loss of efficiency in using a weighted instead of a simple random sample, when the variable of interest is not associated with the weights.

14.4.2 Model-based analysis of data divided into parts

Let us suppose that the research question is how metacognitive competence depends on gender, age (also reflected by grade), socio-economic status, and immigrant status. The following explanatory variables are used:

- * Gender; female = 0, male = 1.
- * Grade, originally ranging from 8 to 12; centered at 10, new range -2 to +2.
- * Age, between 15 and 17 years, centered at 16 years.

- * ESCS, the PISA index of economic, social and cultural status of students; in this data set this has mean 0.15 and standard deviation 0.92.
- * Immigration status, recoded to: 0, at least one parent born in the USA; 1, second generation (born in the USA, both parents born elsewhere); 2: first generation (born outside the USA, parents likewise). This ordinal variable is treated as having a linear effect.

For ESCS and immigration status, the school average is also included in the model, to allow differences between within-school and between-school regression coefficients. These school means are denoted Sch-imm and Sch-ESCS. The centering means that the intercept refers to girls of age 16 and grade 10, with the other variables mentioned taking value 0.

The main design variables are the stratification variables, public/private and region; and school size, determining inclusion probabilities within strata. Of these variables, public/private is of primary interest, the other two variables are of secondary interest only. Therefore, we do not initially include region and school size in the model and rather follow method 3 (p. 226): divide the data set into parts dependent on the design, and apply the hierarchical linear model to each of the parts. Since only 11 out of the 165 schools are private and this is a category of interest, the private schools are considered as one part, and the public schools are divided into four groups according to the school weights (14.19). The private schools have too few level-two units for a good multilevel analysis, but this is an intermediate step only and will give an impression of whether they differ from the public schools. Table 14.4 presents the results. The classes denoted by ‘weight 1’ to ‘weight 4’ are the schools with the smallest to largest weights, therefore the relatively largest to smallest schools. Pupils with missing data on one or more variables were left out; this amounted to about 10% of the data set.

The group of private schools seems to be the most different from the rest, but the sample size is small. The main difference in this respect are the larger effects of the school means of ESCS and of immigrant status; the lower intercept for the private schools may be related to the higher effect of school mean of ESCS together with the fact that these schools on average have a higher ESCS than the public schools. The private schools also have a larger intercept variance, but this is an estimate based on only 11 schools and therefore not very precise. The estimates for the four weight groups of public schools do not differ appreciably, given their standard errors. These differences can be assessed more formally using test (3.42). The variable that leads to the most strongly different estimates is the school mean of immigration status, for which this test applied to the estimates and their standard errors in Table 14.4 leads to the test statistic $C = 7.0$, a chi-squared variable with 3 degrees of freedom, with $p = 0.07$. Being the smallest p -value out of a set of eight variables, this is not really alarming. However, it turns out that also the average immigrant status in the sample is different in the four different weight groups of public schools, ranging from 0.08 in weight group 4 to 0.59 in weight group 1. It is likely that this reflects urbanization rather than weights, or school size, *per se*. Indeed, the average immigrant status in the sample ranges from 0.07 in villages to 0.76 in large cities. Therefore urbanization, a variable in five categories, is added to the data and the estimation is repeated. The results are in Table 14.5. It should be noted that for the private schools this is hardly a meaningful model, as it includes four school variables for 11 schools.

Controlling for urbanization hardly affects the effect of within-school variables and makes the coefficients for the two school averages, especially of immigration status, more

Table 14.4: Estimates for model for metacognitive competence for five parts of the data set.

	Private		Weight 1		Weight 2		Weight 3		Weight 4	
<i>N(schools)</i>	11		38		39		38		39	
<i>Fixed effects</i>	Par.	S.E.	Par.	S.E.	Par.	S.E.	Par.	S.E.	Par.	S.E.
Intercept	-0.72	0.48	-0.12	0.10	-0.15	0.10	-0.17	0.06	-0.18	0.07
Male	-0.27	0.12	-0.34	0.06	-0.33	0.06	-0.28	0.06	-0.35	0.06
Age	-0.05	0.23	-0.05	0.12	-0.18	0.12	-0.26	0.11	-0.01	0.12
Grade	0.25	0.12	0.20	0.06	0.24	0.07	0.28	0.06	0.18	0.07
Immigrant	-0.13	0.09	0.01	0.04	0.05	0.05	-0.03	0.06	0.06	0.09
ESCS	0.04	0.08	0.10	0.04	0.14	0.04	0.08	0.04	0.11	0.04
Sch-imm	0.51	0.37	0.04	0.14	0.14	0.16	0.10	0.16	-0.29	0.36
Sch-ESCS	0.74	0.40	0.16	0.12	0.11	0.12	0.25	0.13	0.10	0.13
<i>Variances</i>	Var.		Var.		Var.		Var.		Var.	
School lev.	0.14		0.04		0.06		0.02		0.05	
Student lev.	0.85		0.93		0.97		0.95		0.95	

similar across the four weight categories of public schools. The parameters of the school averages for the private schools, on the other hand, now deviate even more from those of the public schools than without the control for urbanization.

From this exercise we have learnt that there are differences within the set of public schools, dependent on the design, relating to the between-school effect of immigrant status, and that these differences can be attenuated by controlling for urbanization. Furthermore, we saw that the private schools differ from the public schools with respect to the coefficients of various of the explanatory variables. These conclusions will be utilized in the further analyses.

14.4.3 Inclusion of weights in the model

We now continue with the model as specified in Table 14.5, separately for the data sets of the public schools and of the private schools. The level-one weights (14.18) vary so little that it is meaningless to investigate their effects. Including the weights (14.19) in the data set for the public schools, as main effects, as interactions, or in the random part of the model, does not lead to significant improvements of the model. The same holds for adding school size, also after a square root transformation to make its distribution less skewed. It is not interesting to report the numbers here. This suggests that for constructing a model for the metacognitive competence, given that the model controls for urbanization, it might be possible to get good results without further taking the design into account. In the sequel we shall present the model-based estimates and compare them with the design-based estimates.

Table 14.5: Estimates for model for metacognitive competence, including urbanization, for five parts of the data set.

	Private		Weight 1		Weight 2		Weight 3		Weight 4	
Nschools	11		38		39		38		39	
<i>Fixed effects</i>	Par.	S.E.	Par.	S.E.	Par.	S.E.	Par.	S.E.	Par.	S.E.
Intercept	-0.90	0.40	-0.16	0.12	0.03	0.13	0.01	0.13	-0.01	0.23
Male	-0.27	0.11	-0.33	0.06	-0.33	0.06	-0.28	0.06	-0.35	0.06
Age	-0.07	0.23	-0.04	0.12	-0.19	0.12	-0.27	0.11	-0.01	0.12
Grade	0.26	0.12	0.20	0.06	0.25	0.07	0.28	0.06	0.18	0.07
Immigrant	-0.13	0.09	0.01	0.04	0.05	0.05	-0.03	0.06	0.06	0.09
ESCS	0.03	0.08	0.10	0.04	0.14	0.04	0.08	0.04	0.11	0.04
Sch-imm	0.85	0.33	0.11	0.15	-0.00	0.18	0.11	0.17	0.12	0.43
Sch-ESCS	1.06	0.33	0.19	0.13	0.14	0.13	0.21	0.14	0.15	0.14
Large city	-0.80	0.26	-0.10	0.12	0.01	0.17	-0.32	0.18	-0.51	0.34
Town	-0.36	0.24	—	—	0.12	0.12	-0.17	0.13	-0.38	0.26
Small town	—	—	—	—	0.02	0.21	-0.42	0.17	-0.22	0.13
Village	—	—	—	—	—	—	-0.23	0.17	-0.12	0.24
<i>Variances</i>	Var.		Var.		Var.		Var.		Var.	
School lev.	0.05		0.04		0.05		0.02		0.05	
Student lev.	0.85		0.93		0.97		0.95		0.95	

Reference category for urbanization is ‘city’.

For the private schools a satisfactory analysis is impossible because there are only 11 schools in the sample. Further, the effective sample size as calculated by (14.9a), using the weights (14.19) for only the set of private schools, is only 3.1. This suggests that on the basis of this design it is not possible to say anything about effects of school variables on metacognitive competence for private schools.

14.5 How to assign weights in multilevel models

Design-based analytic inference usually is based on *superpopulation models*. As usual, the survey takes a sample from a finite population, for example, the set of all pupils of a specific school type in a certain country. In addition, it is assumed that this finite population is a sample from a hypothetical infinite population, representing the scientific regularities that are being analyzed, one may say, alternative situations that also ‘could have been the case’.

Thus there are two steps of inference: from the sample to the finite population, and from the finite population to the infinite population. The first step uses the sampling weights to remove the biases that may have been caused by informative sampling designs. The second step is a model-based part of the analysis. In extreme cases the sample covers the entire finite population, and only the second step of inference remains. (Inferential procedures are based, however, on a combined model for the finite population and the superpopulation, and one should not expect two steps to be discernible in the procedures.) This shows why it is meaningful to apply statistical inference to analytic questions even if one has observed the entire (finite) population. It also shows that a superpopulation model leads to larger uncertainty (larger standard errors) than a sampling model with only a finite population: the inference from finite population to superpopulation adds extra uncertainty.

In the multilevel case, the superpopulation has a multilevel structure and the hierarchical linear model holds here; the finite population can be regarded as a multi-stage sample from this superpopulation, so that the hierarchical linear model also holds for the finite population. The observed sample is drawn from the finite population according to a sampling design which may be complex, involving stratification and clustering. If the residuals are correlated with the design variables, a situation called *informative sampling*, the hierarchical linear model does not hold in the same form for the observed sample as it does for the population, and analytic inference has to take the sample design into account to provide unbiased estimators. The preceding sections presented ways to find out whether it seems plausible that the sampling design is noninformative, and proposed to proceed with a regular hierarchical linear model analysis if this is indeed the case. The current section discusses methods of parameter estimation when there is likely to be some association between sampling design and residuals. These methods use the sampling weights. Like the rest of this chapter, the presentation here focuses on two-level models. Level-two units also are called *clusters*.

In the first place it should be mentioned that if there are weights only at level two, and sampling at level one (within the sampled clusters) was done randomly, so that the level-one sampling weights w_{ij} do not depend on i , then the weighting is relatively straightforward, and the problems and ambiguities mentioned below do not occur. Complications occur if there are nonconstant weights at level one.

A second point is that the design-based methods may be unreliable if the number of clusters is small, where the boundary line between small and large enough will normally be between 20 and 40. Multilevel surveys with a small number of units at the highest level, for example, country-comparative studies with rather few countries, in many cases fortunately have large sample sizes at the lower level. In such cases a two-step approach (Section 3.7; Achen, 2005) is preferable to a multilevel design-based estimation for the entire survey: the first step consists of separate analyses for each cluster, the second step of combination of the results, using the estimated clusterwise regression coefficients as dependent variables. If the effective sample sizes per cluster are large, then the parameters can be estimated very precisely for each cluster separately. Then it is not necessary to borrow strength from other highest-level units, so then this argument for conducting a multilevel analysis across clusters does not apply. A practical advantage of this approach is that the within-cluster survey will often have only one level of weights, so that the design-based estimation by cluster can proceed in the more straightforward single-level weighted way (e.g., Fuller, 2009, Chapter 6; Lohr, 2010, Chapter 11).

The weighting methods for hierarchical linear models most often used in the literature are the method developed by Pfeffermann et al. (1998), using the sampling weights in a probability-weighted iterative generalized least squares (PWIGLS) method; and the method developed by Rabe-Hesketh and Skrondal (2006) and also by Asparouhov (2006), extending earlier work by Grilli and Pratesi (2004), maximizing a pseudo-likelihood incorporating the weights. The method of Rabe-Hesketh and Skrondal (2006) and Asparouhov (2006) is applicable to more general multilevel designs (e.g., three or more levels), and for nonnormally distributed dependent variables (generalized linear mixed models) may be expected to perform better than the method of Pfeffermann et al. (1998).

These two weighting methods also differ in the estimates used for the standard errors (or, more generally, the covariance matrices of the estimators). Pfeffermann et al. (1998) use a design-based estimator, while Rabe-Hesketh and Skrondal (2006) and Asparouhov (2006) use a sandwich-type estimator (see Section 12.2). MLwiN, HLM 6, and LISREL 8.7 implement the methods of Pfeffermann et al. (1998), with a choice between the design-based estimator and a sandwich estimator for standard errors. The method of Rabe-Hesketh and Skrondal (2006) and Asparouhov (2006) is implemented in Mplus and Stata (gllamm). A warning is in order when using weights in statistical software: it is always important to determine whether the weighting method uses probability weights (also called design weights) or precision weights (also called frequency weights) (cf. the discussion in Section 14.2). For the purposes of design-based estimation the former type of weights is required. For multilevel models, a telltale sign of the distinction is that probability weighting methods require separate weights (14.8) for the different levels.

In addition to the choice between estimation methods, there is a choice of the scaling method for the level-one weights: no scaling, or the two types of scaling 1 and 2 mentioned above in (14.11).

Thus quite a number of procedures exists, based on the combination of estimation method and scaling method. A clear comparison from a mathematical point of view is given by Bertolet (2008, Section 2.2; 2010).

Properties of these model-based estimators

The following results have been obtained concerning the properties of the PWIGLS method of Pfefferman et al. (1998) and the pseudo-maximum likelihood method of Asparouhov (2006) and Rabe-Hesketh and Skrondal (2006).

The weighted estimators for the fixed parameters according to the method of Asparouhov (2006) and Rabe-Hesketh and Skrondal (2006) are approximately unbiased if the effective cluster sizes are large, for both types of scaling, provided that the level-two random effects are independent of the design variables (Asparouhov, 2006). The estimators for any of these methods are consistent for the fixed effects if the number of clusters becomes large, and for the parameters of the random part if the number of clusters as well as all effective cluster sizes become large (Bertolet, 2008). Simulation studies comparing various specifications for a single estimation method were presented in the original articles by Pfeffermann et al. (1998), Rabe-Hesketh and Skrondal (2006), and Asparouhov (2006).

Bertolet (2008, 2010) presents a careful summary of earlier simulation studies, and presents a simulation study comparing the two estimation methods and various scaling methods.

With respect to the choice of a scaling method, it can be concluded from published theoretical and simulation studies that, if it is likely that the design is informative, some kind of scaling should be used. For large cluster sizes, the differences between the scaling methods are small. However, evidence about the choice between scaling methods 1 and 2 for smaller cluster sizes is equivocal. Based on theoretical arguments for simple models, Potthoff et al. (1992) as well as Korn and Graubard (2003) prefer scaling method 1. Based on simulations, Pfeffermann et al. (1998), Rabe-Hesketh and Skrondal (2006), and Asparouhov (2006) conclude with a tentative preference for scaling method 2. Stapleton (2002) reports simulation studies of multilevel structural equation models where scaling method 1 performs better than method 2. Grilli and Pratesi (2004), on the other hand, found that the procedure of Pfeffermann et al. with scaling method 2 performs well in many realistic situations. Bertolet (2008, 2010) finds in simulations on balance a preference for scaling method 1.

The results are quite encouraging for the estimators of the fixed effects for medium sample sizes, but suggest that coverage probabilities of confidence intervals require large sample sizes. These coverage probabilities, which are equivalent to type I error rates of hypothesis tests, are essential criteria for the reliability of these methods and integrate bias, estimation variance, and standard error estimation, but unfortunately many published simulation studies fail to report them. In a simple linear two-level model with 35 clusters, Pfeffermann et al. (1998) report reasonable precision for cluster sizes of about 40, but do not report coverage probabilities. Rabe-Hesketh and Skrondal (2006) report good coverage probabilities for a simulation of a two-level logistic regression model of 500 clusters with cluster size 50. Asparouhov (2006) presents a simulation study of an empty linear model with 100 clusters, and finds coverage probabilities for the fixed effect which are sometimes good and sometimes (depending on intraclass correlations and degree of informativeness of the sampling design) much too low for cluster size 20, and mostly good coverage probabilities for cluster size 100. Bertolet (2008, 2010) finds unacceptably low coverage probabilities for linear models and normal distributions in simulations of 35 clusters with cluster size 20. It must be concluded that published simulations have only covered a few combinations of sample sizes and model specifications, that tests and confidence intervals can be unreliable for small sample sizes, and that as yet there is no support for good coverage probabilities unless there are at least 100 clusters and cluster sizes are at least 50.

The various methods have also been compared for empirical data sets. This was also done in the original articles introducing the methods. Further, Carle (2009) compares all these methods for a data set on children with special health care needs in the USA. Zaccarin and Donati (2009) compare different weighting and scaling methods for a subset of the 2003 PISA data, an international comparative study of educational outcomes. Carle finds only quite small differences between different methods for his data set, but Zaccarin and Donati find large differences in cases where the level-one weights are strongly variable. Neither of these studies, however, attempted to achieve extensive control for relevant level-one predictor variables that might have explained away the association between the outcome variable and the sampling design.

Conclusions

All these simulation studies are based on limited designs and cannot be regarded as conclusive. It can be concluded that for the estimation of fixed effects in linear multilevel models, the methods mentioned all give good results, perhaps excluding the case of a very low number of clusters. For the estimation of random effects and nonlinear multilevel models, the method of Rabe-Hesketh and Skrondal (2006) and Asparouhov (2006) seems better than the method of Pfeffermann et al. (1998). It is necessary to apply scaling to the level-one weights, but it is not clear which of the two methods (14.11a) or (14.11b) is better, and this may depend on the design, the parameter values, and what is the parameter of main interest. Carle (2009) suggests using both scaling methods; and, if these produce different results, to conduct simulation studies to determine, for the given design and plausible parameter values, which is the preferable scaling method. This seems to be good advice.

The estimation of standard errors, however, still is a problem. Coverage probabilities of confidence intervals (and therefore also type I error rates of hypothesis tests) can be too low for small sample sizes. Extant simulation studies suggest that a survey of 35 clusters of size 20 is too small to apply the design-based methods discussed here (Bertotet, 2008, 2010), while 100 clusters of size 100 is sufficient (Asparouhov, 2006). How generalizable this is, and what happens between these two boundaries, is still unknown. Perhaps bootstrap standard errors (see below) give better results, but this still is largely unexplored.

For two-level surveys with few clusters, but large cluster sizes, a good alternative may be to follow a two-step approach with a first step of separate estimations for each cluster, where single-level weighting may be used to account for sampling weights (cf. Skinner, 1989; Lohr, 2010, Chapter 11) and a second step combining the results for the different clusters (cf. Section 3.7).

Example 14.1 Design-based and model-based analysis of PISA data.

We continue the example of Section 14.4 of metacognitive competence in the PISA 2009 data for the USA. In Section 14.4 we learned that level-two weights are important in this data set; schools were sampled with inclusion probabilities dependent on school size. Level-one weights are hardly variable, and therefore not important. We also learned that it may be good to control for urbanization. Furthermore, private schools were perhaps different from public schools; but the number of private schools sampled, 11, was too low to say much about this question with any confidence.

Based on this information, it was decided first to retain the entire data set, control for main effects of urbanization as well as private/public; and compare the design-based estimates, computed by MLwiN according to the method of Pfeffermann et al. (1998), employing scaling method 2, with the model-based estimates according to the hierarchical linear model. The results are presented in Table 14.6. Deviances are not given, because these are not meaningful for the pseudo-likelihood method. To compare the two kinds of estimate, Asparouhov's (2006) informativeness measure I_2 is given, as presented in (14.17).

The results for the student-level effects correspond rather well, but those for the school-level effects do not. I_2 expresses the difference in terms of the model-based standard error. Values larger than 2 can be considered unacceptable, and these occur for the two school averages and also for some of the control variables for urbanization.

To diagnose the differences, first the private schools were left out, because of the possibility of differences between private and public schools and the lack of enough information on the private schools, and the two estimates were compared again. The differences remained (results not shown here). Therefore a residual analysis was done (Chapter 10), with particular attention to the school sizes, as these were the main determinants of the inclusion probabilities. It turned out that there is

Table 14.6: Design-based and model-based estimates for model for metacognitive competence, entire data set.

	Design-based		Model-based		
<i>Fixed effects</i>	Par.	S.E.	Par.	S.E.	I_2
Intercept	-0.077	0.089	-0.075	0.058	0.03
Male	-0.350	0.040	-0.321	0.029	1.00
Age	-0.134	0.102	-0.121	0.055	0.24
Grade	0.192	0.043	0.224	0.031	1.03
Immigrant	-0.014	0.031	0.008	0.027	0.81
ESCS	0.075	0.035	0.107	0.019	1.68
Sch-imm	0.240	0.104	0.094	0.068	2.41
Sch-ESCS	0.426	0.119	0.189	0.049	4.84
Private	-0.211	0.179	-0.008	0.118	1.72
Large city	-0.406	0.121	-0.152	0.081	3.14
Town	-0.390	0.111	0.093	0.057	8.47
Small town	-0.382	0.108	0.172	0.072	7.69
Village	-0.203	0.110	-0.087	0.073	1.59
<i>Variances</i>	Var.		Var.		
School lev.	0.031	0.007	0.040	0.009	
Student lev.	0.893	0.083	0.941	0.018	

I_2 is Asparouhov's (2006) informativeness measure (14.17).

one outlying school, having 6,694 pupils while the other school sizes ranged from 100 to 3,592, and with results concerning metacognitive competence quite different from the other schools. This school was also left out of the data set, leaving a total sample of 153 public schools with 4,316 pupils. (An alternative would be to retain this school in the data set and represent it by a dummy variable.) In addition, school size was used as an additional control variable; in view of the skewness of school size, the square root of school size was used, centered at 35 (corresponding to a school size of 1,225, close to average school size). Table 14.7 presents the results of the design-based and model-based estimates for this smaller data set.

The results of the two approaches now are more in line with each other. The main difference now is the coefficient of age, which is just significant for the model-based analysis, and small but still positive for the design-based analysis, and has an informativeness measure slightly above 2. Further explorations led to the conclusion that this difference may have to do with an interaction between age and school size, or urbanization – these variables are too strongly associated to disentangle their interaction effects. Table 14.8 shows the results of the two types of analysis, now including the interaction of age with school size.

It turns out that in this model the differences between the two approaches are further reduced, with all informativeness measures less than 2, and identical conclusions with respect to significance of the variables in the original model of interest. The remaining differences may be regarded as random

Table 14.7: Design-based and model-based estimates
for model for metacognitive competence, public schools without outlier.

	Design-based		Model-based		
<i>Fixed effects</i>	Par.	S.E.	Par.	S.E.	I_2
Intercept	-0.050	0.068	-0.122	0.061	1.18
Male	-0.359	0.046	-0.318	0.032	1.28
Age	0.011	0.080	-0.117	0.059	2.17
Grade	0.162	0.040	0.215	0.032	1.66
Immigrant	0.019	0.034	0.017	0.030	0.07
ESCS	0.090	0.032	0.111	0.019	1.11
Sch-imm	0.114	0.092	0.068	0.078	0.59
Sch-ESCS	0.204	0.081	0.143	0.054	1.13
$\sqrt{\text{school size}} - 35$	0.004	0.005	0.004	0.003	0.00
Large city	-0.193	0.091	-0.094	0.078	1.27
Town	-0.094	0.059	-0.032	0.060	1.03
Small town	-0.128	0.088	-0.100	0.080	0.35
Village	0.034	0.102	0.023	0.088	0.13
<i>Variances</i>	Var.		Var.		
School lev.	0.035	0.010	0.038	0.009	
Student lev.	0.939	0.085	0.940	0.018	

I_2 is Asparouhov's (2006) informativeness measure (14.17).

variation. The model-based results have smaller standard errors, confirming the greater precision of these estimators.

As a final conclusion of this data analysis, it seems reasonable to present the model-based results of Table 14.8. The main interpretations are the following:

- * Girls have on average higher metacognitive competences than boys.
- * Metacognitive competences increase with grade level.
- * Within the same grade, older students have on average lower metacognitive competence than younger ones. This may be because of reasons that also led to delays in the school careers of the older pupils in the same grade.
- * Students from families with more economic, social and cultural resources, and those in schools where on average families have more of these resources, have higher metacognitive competences.
- * Unexplained variation is mainly at the individual level, and quite small at the school level.

This should be accompanied, however, by the following caveats:

- * The observational and cross-sectional nature of the data precludes causal interpretations of the above conclusions.

Table 14.8: Design-based and model-based estimates for model for metacognitive competence, public schools without outlier, with more extensive controls.

Fixed effects	Design-based		Model-based		
	Par.	S.E.	Par.	S.E.	I_2
Intercept	-0.067	0.065	-0.120	0.061	0.87
Male	-0.357	0.045	-0.319	0.032	1.19
Age	-0.083	0.065	-0.113	0.058	0.52
Grade	0.167	0.040	0.216	0.032	1.53
Immigrant	0.019	0.034	0.017	0.030	0.07
ESCS	0.090	0.032	0.111	0.019	1.11
Sch-imm	0.108	0.092	0.065	0.060	0.72
Sch-ESCS	0.203	0.081	0.143	0.054	1.11
$\sqrt{\text{school size}} - 35$	0.002	0.005	0.003	0.003	0.33
Large city	-0.194	0.091	-0.093	0.078	1.29
Town	-0.096	0.060	-0.033	0.060	1.05
Small town	-0.134	0.089	-0.101	0.080	0.41
Village	0.036	0.104	0.023	0.088	0.15
Age $\times \sqrt{\text{school size}} - 35$	-0.011	0.005	-0.004	0.004	1.75
Variances		Var.	Var.		
School lev.	0.036	0.010	0.038	0.009	
Student lev.	0.937	0.084	0.940	0.018	

I_2 is Asparouhov's (2006) informativeness measure (14.17).

- * This represents the public schools only. The private schools seem to present a somewhat different picture, but their sampled number is too small to draw conclusions about these differences.
- * There was one outlying very large public school which also seemed to present a different picture, and which was left out of the data for the final results.
- * There may be interactions of age with other variables on metacognitive competence. Note that the effect of age is already controlled for grade, and the age range in this data is small (between 15.2 and 16.4 years). These interactions could be with school size or with urbanization. These variables are associated and their interaction effect cannot be disentangled.
- * A further analysis could investigate differences between the four regions used for the stratification. These differences were not considered here.

Other methods

Since this is an area in development, it may be good to mention some other procedures that have been proposed in the literature.

Korn and Graubard (2003) proposed another weighting method, requiring knowledge of higher-order inclusion probabilities which often is not available; therefore it is less widely applied. However, in their simulations their method does appear to perform well. Bertolet (2008, Section 2.2; 2010) explains how their method compares to the two methods mentioned above.

Estimation of standard errors by bootstrap methods was proposed and studied by Grilli and Pratesi (2004) and by Kovačević et al. (2006), and seems to perform well. Given the difficulties for standard error estimation, this seems a promising method.

Pfefferman et al. (2006) propose a totally different approach. They construct a simultaneous model for the random sample (the ‘inclusion events’) and the dependent variable, conditional on inclusion in the sample. This approach does not use weighting: it is model-based, not design-based.

Little (2004) and Gelman (2007) suggest a Bayesian approach by giving random effects to subsets of the population having constant (or similar) inclusion probabilities. This approach sacrifices unbiasedness, but may obtain smaller posterior variances or smaller mean squared errors in return.

14.6 Appendix. Matrix expressions for the single-level estimators

For readers with a knowledge of matrix algebra it may be helpful to see the matrix expressions for the PWE and WLS estimators. The linear model is defined by

$$Y = X\beta + R.$$

In both cases, (14.1) as well as (14.4), the estimator is defined by

$$\hat{\beta}^W = (X'WX)^{-1}X'WY.$$

In the first case, (14.1), the most frequently made assumption is that the residuals are homoscedastic, $\text{cov}(R) = \sigma^2 I$, giving the covariance matrix

$$\text{cov}(\hat{\beta}^W) = \sigma^2(X'WX)^{-1}(X'W^2X)(X'WX)^{-1}. \quad (\text{design weights}) \quad (14.20)$$

In the second case (14.4) it is assumed that the residuals are heteroscedastic, $\text{cov}(R) = \sigma^2 W^{-1}$, yielding

$$\text{cov}(\hat{\beta}^W) = \sigma^2(X'WX)^{-1}. \quad (\text{precision weights}) \quad (14.21)$$

14.7 Glommary

Two-stage sample. A sample design where level-two units are chosen independently with some probability, and given that a level-two unit j has been selected, a sample of level-one units within this level-two unit is chosen.

Cluster sample. A sample design where level-two units are chosen independently with some probability, and given that a level-two unit j has been selected, all level-one units within this level-two unit are included in the sample.

Stratified sample. A sample design where the population is divided into subsets (strata), and samples are drawn independently within each stratum, with sampling proportions which may differ between strata.

Inclusion probabilities, sampling probabilities. At level two, the probabilities that given level-two units are included in the sample; at level one, the probabilities that given level-one units are included in the sample, under the condition that the level-two unit of which they are a part is already included.

Design variables. Variables that determine the inclusion probabilities and other elements of the design.

Descriptive inference. The estimation and testing of descriptive parameters of the population, as done in official statistics.

Analytic inference. Statistical inference about the way in which a variable depends on, or is associated with, other variables.

Model-based inference. Statistical inference based on the assumption that the data can be regarded as the outcome of a probability model (e.g., a linear model with normal residuals) with some unknown parameters. The inference is directed at obtaining knowledge about the likely values of these parameters.

Design-based inference. Statistical inference based on the probability mechanism used for the sample selection. Somewhat more generally, a *design-based approach to analytic inference* refers to analyzing the survey data according to a postulated model, in this book a hierarchical linear model, while taking the design into account.

Sampling weights or survey weights. Weights for data points, inversely reflecting the probability that any particular population element is included in the sample.

Precision weights, analytical weights, frequency weights. Weights for data points inversely reflecting the residual standard deviations. An example, the case of frequency weights, is the case where a data point with a larger weight is the average of a larger number of basic population elements.

Probability weighted estimator. An estimator obtained by using weights that are reciprocals of inclusion probabilities. If appropriately used, this gives unbiased or nearly unbiased estimators.

Weighted least squares. An estimation method defined by minimizing the sum of weighted squared deviations between the observed values of the dependent variable and their fitted values.

Effective sample size. A sample, with a complex sample design and having effective sample size n_{eff} , gives the same amount of information as a simple random sample of size n_{eff} ; the usual formula holds, however, only for the estimation of population means of homoscedastic random variables.

Design effect. Effective sample size divided by actual sample size.

Model of interest. The probability model on which the research is substantively focused; in this book, it takes the form of a hierarchical linear model.

Informative design. A design is noninformative for a model of interest if the residuals in the model of interest are independent of the design variables.

Scaling of weights. Methods to scale the level-one weights within a given level-two unit, either (method 1) to make them sum to the effective sample size, or (method 2) to make them sum to the actual sample size.

Superpopulation model. The assumption of an (infinite) superpopulation for which the hierarchical linear model holds; and from which the actual data are a sample, drawn perhaps according to a complex sample design.

Pseudo-maximum likelihood. An approach to estimation in superpopulation models, based on a modified maximum likelihood equation in which weights are inserted that reflect the sample design.

Sandwich estimator or robust estimator. A procedure for estimating standard errors for which the mathematical equation has the form of a sandwich (one matrix sandwiched between two other matrices), and which is more robust against deviations from the model than standard errors calculated directly from the model equations.



15

Longitudinal Data

The paradigmatic nesting structure used up to now in this book has been the structure of individuals (level-one units) nested in groups (level-two units). The hierarchical linear model is also very useful for analyzing repeated measures, or longitudinal data, where multiple repeated measurements of the same variable are available for a sample of individual subjects. Since the appearance of the path-breaking paper by Laird and Ware (1982), this is the main type of application of the hierarchical linear model in the biological and medical sciences. This chapter is devoted to the specific two-level models and modeling considerations that are relevant for data where the level-one units are measurement occasions and the level-two units individual subjects.

Of the advantages of the hierarchical linear model approach to repeated measures, one deserves to be mentioned here. It is the flexibility to deal with unbalanced data structures, for example, repeated measures data with fixed measurement occasions where the data for some (or all) individuals are incomplete, or longitudinal data where some or even all individuals are measured at different sets of time points. Here it must be assumed that the fact that a part of the data was not observed does not itself contain some information about the unobserved values; see Chapter 9 for a treatment of incomplete data.

The topic of repeated measures analysis is too vast to be treated in one chapter. For a general treatment of this topic we refer the reader to, for example, Maxwell and Delaney (2004) and Fitzmaurice et al. (2004). In economics this type of model is discussed mainly under the heading of panel data; see, for example, Baltagi (2008), Chow (1984), and Hsiao (1995). Some textbooks treating repeated measures specifically from the perspective of multilevel modeling are Verbeke and Molenberghs (2000), Singer and Willett (2003, Part I), and Hedeker and Gibbons (2006). The current chapter only explains the basic hierarchical linear model formulation of models for repeated measures.

This chapter is about the two-level structure of measurements within individuals. When the individuals, in their turn, are nested in groups, the data have a three-level structure: longitudinal measurements nested within individuals nested within groups. Models for such data structures can be obtained by adding the group level as a third level to the models of this chapter. Such three-level models are not explicitly treated in this chapter. However, this three-level extension of the fully multivariate model of Section 15.1.3 is the same as the multivariate multilevel model of Chapter 16.

The chapter is divided into two parts. The first part deals with fixed occasion data structures, defined by a fixed set of measurement occasions for all individual subjects – where any number of measurements may, however, be randomly missing for any of the subjects. For these data structures we treat several specifications of the residual covariance patterns, representable by a random intercept and random slopes of transformed time variables. The most encompassing specification is the fully multivariate model, which leaves the residual covariance matrix completely free. This model can also be used for multivariate data that do not have a longitudinal nature, that is, for multivariate regression analysis.

The second part of the chapter is about variable occasion designs, where the number and timing of measurement occasions is unrestricted across subjects. The measurements here must be arranged along an underlying continuum such as time or age, and may be regarded as measurements of a population of curves. Polynomial, piecewise linear, and spline functions are treated as families of curves that may be useful to represent such populations. These curves may then be explained by individual (level-two) variables, by interactions between individual variables and transformed time variables, and by changing covariates. Finally, some remarks are made about the possibility of replacing the assumption of independent level-one residuals by weaker assumptions under which the level-one residuals have some kind of autocorrelation structure.

15.1 Fixed occasions

In fixed occasion designs, there is a fixed set $t = 1, \dots, m$ of measurement occasions. For example, in a study of some educational or therapeutic program, we might have an intake test, pretest, mid-program test, post-test, and follow-up test ($m = 5$). Another example is a study of attitude change in early adulthood, with attitudes measured shortly after each birthday from the 18th to the 25th year of age. If data are complete, each individual has provided information on all these occasions. It is quite common, however, that data are incomplete. When they are, we assume that the absent data are missing at random, and the fact that they are missing does not itself provide relevant information about the phenomena studied (cf. Chapter 9).

The measurement occasions are denoted by $t = 1, \dots, m$, but each individual may have a smaller number of measurements because of missing data. Y_{ti} denotes the measurement for individual i at occasion t . It is allowed that, for any individual, some measurements are missing. Even individuals with only one measurement do not need to be deleted from the data set: they do contribute to the estimation of the between-individual variance, although they do not give information about within-individual variability.

Note that, differently from the other chapters, the level-one units are now the measurement occasions indexed by t , while the level-two units are the individuals indexed by i .

The models treated in this section differ primarily with respect to the random part. Choosing between them amounts to specifying the random part of the model, and there are three kinds of motive that should be taken into account for this purpose.

The first motive is that if one works with a random part that is not an adequate description of the dependence between the m measurements (i.e., it does not satisfactorily represent the covariance matrix of these measurements), then the standard errors for estimated coefficients in the fixed part are not reliable. Hence, the tests for these coefficients also are unreliable. This motive points to the least restrictive model, that is, the fully multivariate model, as the preferred one.

The second motive is the interpretation of the random part. Simpler models often allow a nicer and easier interpretation than more complicated models. This point of view favors the more restrictive models, for example, the random intercept (compound symmetry) and random slope models.

The third motive is the possibility of getting results at all and the desirability of standard errors that are not unnecessarily large (while not having a bias at the same time). If the amount of data is relatively small, the estimation algorithms may not converge for models that have a large number of parameters, such as the fully multivariate model. Or even if the algorithm does converge, having a bad ‘data-to-parameter ratio’ may lead to large standard errors. This suggests that, if one has a relatively small data set, one should not entertain models with too many parameters.

To conclude, one normally should use the simplest model for the random part (i.e., the model having the smallest number of parameters and being the most restrictive) that still yields a good fit to the data. The model for the random part can be selected with the aid of deviance tests (see Section 6.2).

15.1.1 The compound symmetry model

The classical model for repeated measures, called the compound symmetry model (e.g., Maxwell and Delaney, 2004), is the same as the random intercept model. It is sometimes referred to as the random effects model or the mixed model, but one should realize that it is only a simple specification of the many possibilities of the random effects model or mixed model. If there are no explanatory variables except for the measurement occasions, that is, the design is a pure *within-subjects design*, the expected value for measurement occasion t can be denoted by μ_t and this model can be expressed by

$$Y_{ti} = \mu_t + U_{0i} + R_{ti}. \quad (15.1)$$

The usual assumptions are made: the U_{0i} and R_{ti} are independent normally distributed random variables with expectations 0 and variances τ_0^2 for U_{0i} and σ^2 for R_{ti} .

To fit this model, note that the fixed part does not contain a constant term, but is based on m dummies for the m measurement occasions. This can be expressed by the following formula. Let d_{hti} be m dummy variables, defined for $h = 1, \dots, m$ by

$$d_{hti} = \begin{cases} 1 & (t = h), \\ 0 & (t \neq h). \end{cases} \quad (15.2)$$

Then the fixed part μ_t in (15.1) can be written as

$$\mu_t = \sum_{h=1}^m \mu_h d_{hti}$$

and the compound symmetry model can be formulated as

$$Y_{ti} = \sum_{h=1}^m \mu_h d_{hti} + U_{0i} + R_{ti}, \quad (15.3)$$

the usual form of the fixed part of a hierarchical linear model.

Example 15.1 Life satisfaction.

The German Socio-Economic Panel (SOEP) is a longitudinal panel data set for the population in Germany. We use the question on life satisfaction, which has answer categories ranging from 0 (totally unhappy) to 10 (totally happy), and study how it evolves during the ages of 55–60 years, and how it is influenced by retirement. The survey was held each year. We use data collected in 1984–2006 for a total of 1,236 individuals, each of whom participated from 1 to 6 years in this part of the panel study. This is part of a sample of individuals representative of the population of Germany in the year 1984. Individuals were included if they were heads of household, aged 55 years at one of the moments of data collection, and had information about their employment status.

We use six measurement occasions: for ages 55, 56, ..., 60. Thus, the time variable t assumes the values 1 (age 55) through 6 (age 60). The number of observations available for the six ages decreased from 1,224 at $t = 1$ to 799 at $t = 6$. The first number is less than 1,236 because a few observations with missing data were dropped. The nonresponse or nonavailability at various moments may be considered to be independent of the variables being investigated here, so missingness may be considered to be at random.

We begin by considering two models with a random intercept only, that is, with a compound symmetry structure for the covariance matrix. The first is the empty model of Chapter 4. In this model it is assumed that the six measurement occasions have the same population mean. The second model is model (15.1), which allows the means to vary freely over time. In this case, writing out (15.3) results in the model

$$Y_{ti} = \mu_1 d_{1ti} + \mu_2 d_{2ti} + \mu_3 d_{3ti} + \mu_4 d_{4ti} + \mu_5 d_{5ti} + \mu_6 d_{6ti} + U_{0i} + R_{ti}.$$

Applying definition (15.2) implies that the expected score for an individual, for example, at time $t = 2$ (age 56 years), is

$$\hat{Y}_{ti} = \mu_1 \times 0 + \mu_2 \times 1 + \mu_3 \times 0 + \mu_4 \times 0 + \mu_5 \times 0 + \mu_6 \times 0 = \mu_2.$$

The results are in Table 15.1. However, it is dangerous to trust results for the compound symmetry model before its assumptions have been tested, because standard errors of fixed effects may be incorrect if one uses a model with a random part that has an unsatisfactory fit. Later we will fit more complicated models and show how the assumption of compound symmetry can be tested.

The results suggest that individual (level-two) variation is a little more important than random differences between measurement occasions (level-one variation). Further, the means change only slightly from time $t = 1$ to time $t = 6$. The deviance test for the difference in mean between the six time points is borderline significant: $\chi^2 = 21,791.34 - 21,780.70 = 10.64$, $df = 5$; $p = 0.06$. This near-significance is not meaningful in this case, in view of the large data set.

Just as in our treatment of the random intercept model in Chapter 4, any number of relevant explanatory variables can be included in the fixed part. Often there are individual-dependent explanatory variables Z_k , $k = 1, \dots, q$, for example, traits or background characteristics. If these variables are categorical, they can be represented by dummy variables. Such individual-dependent variables are level-two variables in the multilevel

Table 15.1: Estimates for random intercept models.

Fixed effect	Model 1		Model 2	
	Coefficient	S.E.	Coefficient	S.E.
μ_1 Mean at age 55	6.937	0.044	6.882	0.053
μ_2 Mean at age 56	6.937	0.044	6.956	0.054
μ_3 Mean at age 57	6.937	0.044	7.021	0.056
μ_4 Mean at age 58	6.937	0.044	6.907	0.057
μ_5 Mean at age 59	6.937	0.044	6.894	0.059
μ_6 Mean at age 60	6.937	0.044	6.985	0.060
Random effect	Parameter	S.E.	Parameter	S.E.
<i>Level-two (i.e., individual) variance:</i>				
$\tau_0^2 = \text{var}(U_{0i})$	1.994	0.095	1.991	0.094
<i>Level-one (i.e., occasion) variance:</i>				
$\sigma^2 = \text{var}(R_{ti})$	1.455	0.030	1.452	0.030
Deviance	21,791.34		21,780.70	

approach, and they are called *between-subjects variables* in the terminology of repeated measures analysis. In addition, there are often one or more numerical variables describing the measurement occasion. Denote such a variable by $s(t)$; this may, for example, be a measure of the time elapsing between the occasions, or the rank order of the occasion. Such a variable is called a *within-subjects variable*. In addition to the between-subjects and within-subjects variables having main effects, they may have interaction effects with each other. These within–between interactions are a kind of cross-level interaction in the multi-level terminology and represent, for example, differences between individuals in their rate and pattern of change in the dependent variable. Substantive interest in repeated measures analysis often focuses on these interactions.

When the main effect parameter of the between-subjects variable z_{ki} is denoted α_k and the interaction effect parameter between z_{ki} and $s(t)$ is denoted γ_k , the model is given by

$$Y_{ti} = \sum_{k=1}^q \alpha_k z_{ki} + \sum_{k=1}^q \gamma_k z_{ki} s(t) + \sum_{h=1}^m \mu_h d_{hti} + U_{0i} + R_{ti}. \quad (15.4)$$

The fixed part is an extension of (15.1) or the equivalent form (15.3), but the random part is still the same. Therefore this is still called a compound symmetry model. Inclusion in the fixed part of interactions between individual-level explanatory variables and time variables such as $s(t)$ suggests, however, that the random part could also contain random slopes of these time variables. Therefore we defer giving an example of the fixed part (15.4) until after the treatment of such a random slope.

Classical analyses of variance methods are available to estimate and test parameters of the compound symmetry model if all data are complete (e.g., Maxwell and Delaney, 2004). The hierarchical linear model formulation of this model, and the algorithms and software available, also permit the statistical evaluation of this model for incomplete data without additional complications.

Covariance matrix

In the fixed occasion design one can talk about the complete data vector

$$Y^c = \begin{pmatrix} Y_{1t} \\ \vdots \\ Y_{mt} \end{pmatrix}.$$

Even if there were no subject at all with complete data, the complete data vector would still make sense from a conceptual point of view.

The compound symmetry model (15.1), (15.3) or (15.4) implies that for the complete data vector, all variances are equal and also all covariances are equal. The expression for the covariance matrix of the complete data vector, conditional on the explanatory variables, is the $m \times m$ matrix

$$\Sigma(Y^c) = \begin{pmatrix} \tau_0^2 + \sigma^2 & \tau_0^2 & \tau_0^2 & \tau_0^2 \\ \tau_0^2 & \tau_0^2 + \sigma^2 & \tau_0^2 & \tau_0^2 \\ \tau_0^2 & \tau_0^2 & \tau_0^2 & \tau_0^2 \\ \vdots & \vdots & \vdots & \vdots \\ \tau_0^2 & \tau_0^2 & \tau_0^2 & \tau_0^2 + \sigma^2 \end{pmatrix}; \quad (15.5)$$

cf. p. 49. This matrix is referred to as the compound symmetry covariance matrix. In it, all residual variances are the same and all residual within-subject correlations are equal to

$$\rho_I = \rho\{Y_{ti}, Y_{si}\} = \frac{\tau_0^2}{\tau_0^2 + \sigma^2}, \quad t \neq s, \quad (15.6)$$

the residual intraclass correlation which we encountered in Chapter 4.

The compound symmetry model is a very restrictive model, and often an unlikely one. For example, if measurements are ordered in time, the correlation is often larger between nearby measurements than between measurements that are far apart. (Only for $m = 2$ is this condition not so very restrictive. In this case, formula (15.5) only means that the two measurements have the same variance and a positive correlation.)

Example 15.2 Covariance matrix for life satisfaction.

In Table 15.1 for the empty model the estimates $\hat{\sigma}^2 = 1.455$ and $\tau_0^2 = 1.994$ were obtained. This implies that the estimated variances of the observations are $1.455 + 1.994 = 3.449$ and the estimated within-subjects correlation is $\hat{\rho}_I = 1.994/3.449 = 0.58$. This number, however, is conditional on the validity of the compound symmetry model; we shall see below that the compound symmetry model gives a reasonable description of the covariance matrix, although more complicated models do give a significantly better fit.

15.1.2 Random slopes

There are various ways in which the assumption of compound symmetry (which states that the variance of the observations is constant over time and that the correlation between observations is independent of how far apart they are, as is expressed by (15.5)) can be relaxed. In the hierarchical linear model framework, the simplest way is to include one or more random slopes in the model. This makes sense if there is some meaningful dimension, such as time or age, underlying the measurement occasions. It will be assumed here that the index t used to denote the measurement occasions is a meaningful numerical variable, and that it is relevant to consider the regression of Y on t . This variable will be referred to as ‘time’. It is easy to modify the notation so that some other numerical function of the measurement occasion t gets a random slope. It is assumed further that there is some meaningful reference value for t , denoted by t_0 . This could refer, for example, to one of the time points, such as the first. The choice of t_0 affects only the parameter interpretation, not the fit of the model.

Since the focus is now on the random rather than on the fixed part, the precise formulation of the fixed part is left out of the formulas.

The model with a random intercept and a random slope for t is given by

$$Y_{ti} = \text{fixed part} + U_{0i} + U_{1i}(t - t_0) + R_{ti}. \quad (15.7)$$

This model means that the rates of increase have a random, individual-dependent component U_{1i} , in addition to the individual-dependent random deviations U_{0i} which affect all values Y_{ti} in the same way. The random effect of time can also be described as a random time-by-individual interaction.

The value t_0 is subtracted from t in order to allow the intercept variance to refer not to the (possibly meaningless) value $t = 0$ but to the reference point $t = t_0$; cf. p. 76. The variables (U_{0i} , U_{1i}) are assumed to have a joint bivariate normal distribution with expectations 0, variances τ_0^2 and τ_1^2 , and covariance τ_{01} .

The variances and covariances of the measurements Y_{ti} , conditional on the explanatory variables, are now given by

$$\begin{aligned} \text{var}(Y_{ti}) &= \tau_0^2 + 2\tau_{01}(t - t_0) + \tau_1^2(t - t_0)^2 + \sigma^2, \\ \text{cov}(Y_{ti}, Y_{si}) &= \tau_0^2 + \tau_{01}\{(t - t_0) + (s - t_0)\} + \tau_1^2(t - t_0)(s - t_0), \end{aligned} \quad (15.8)$$

where $t \neq s$; we saw the same formulas in (5.5) and (5.6). These formulas express the fact that the variances and covariances of the outcome variables are variable over time. This is called heteroscedasticity. Differentiating with respect to t and equating the derivative to 0 shows that the variance is minimal at $t = t_0 - \tau_{01}/\tau_1^2$ (if t were allowed to assume any value). Furthermore, the correlation between different measurements depends on their spacing (as well as on their position).

Extensions to more than one random slope are obvious; for example, a second random slope could be given to the squared value $(t - t_0)^2$. In this way, one can perform a *polynomial trend analysis* to improve the fit of the random part. This means that one fits random slopes for a number of powers of $(t - t_0)$ to obtain a model that has a good fit to the data and where unexplained differences between individuals are represented as random individual-dependent regressions of Y on $(t - t_0)$, $(t - t_0)^2$, $(t - t_0)^3$, etc. Functions other than polynomials can also be used, for example, splines (see Section 15.2.2). Polynomial trend

analysis is also discussed in Singer and Willett (2003, Section 6.3) and Maas and Snijders (2003).

Example 15.3 Random slope of time in life satisfaction.

We continue the earlier example of life satisfaction for 55–60-year-olds. Note that the observations are spaced a year apart, and a natural variable for the occasions is $s(t) = t$ as used above, that is, age recoded so that $t = 1$ is 55 years and $t = 6$ is 60 years. Thus, the time dimension here corresponds to age. A random slope of age is now added (Model 3). The reference value for the time dimension, denoted by t_0 in formula (15.7), is taken as $t_0 = 1$, corresponding to 55 years of age.

We now investigate whether part of the differences between individuals in life satisfaction and in the rate of change can be explained by birth cohort. Years of birth here range from 1929 to 1951. This variable is centered at the mean, which is 1940. To avoid very small regression coefficients the variable is divided by 10, leading to a variable Z ranging from -1.1 to $+1.1$, used as a between-subjects variable. A difference of one unit in Z corresponds to a difference of 10 years in birth year. The effect of cohort on the rate of change, that is, the interaction effect of birth year with age, is represented by the product of Z with t . The resulting model is an extension of a model of the type of (15.4) with a random slope:

$$Y_{ti} = \mu_t + \alpha z_i + \gamma z_i(t - t_0) + U_{0i} + U_{1i}(t - t_0) + R_{ti}. \quad (15.9)$$

Parameter α is the main effect for birth year, while γ is the interaction effect between birth year and age.

Table 15.2: Estimates for random slope models.

Fixed effect	Model 3		Model 4	
	Coefficient	S.E.	Coefficient	S.E.
μ_1 Effect of age 55	6.883	0.055	6.842	0.055
μ_2 Effect of age 56	6.955	0.055	6.914	0.055
μ_3 Effect of age 57	7.022	0.055	6.988	0.055
μ_4 Effect of age 58	6.912	0.056	6.886	0.057
μ_5 Effect of age 59	6.906	0.058	6.892	0.060
μ_6 Effect of age 60	7.004	0.060	7.005	0.064
α Main effect birth year			-0.394	0.078
γ Interaction birth year \times age			0.049	0.019
Random effect	Parameter	S.E.	Parameter	S.E.
<i>Level-two variation:</i>				
τ_0^2 Intercept variance	2.311	0.126	2.253	0.124
τ_1^2 Slope variance age	0.025	0.005	0.025	0.005
τ_{01} Intercept-slope covariance	-0.103	0.021	-0.097	0.021
<i>Level-one (i.e., occasion) variance:</i>				
σ^2 Residual variance	1.372	0.031	1.371	0.031
Deviance	21,748.03		21,722.45	

The results are given in Table 15.2. To allow deviance tests, all results were calculated using the maximum likelihood estimation procedure. Comparing the deviances of Models 2 and 3 shows that the random slope of age is significant: $\chi^2 = 32.7$, $df = 2$, $p < 0.0001$. This implies that there is a significant deviation from the compound symmetry model, Model 2 of Table 15.1. However, the differences between individuals in rate of change are still rather small, with an estimated inter-individual standard deviation of $\sqrt{0.025} = 0.16$. One standard deviation difference between individuals with respect to slope for age will add up for the age range of 5 years to a difference of $0.16 \times 5 = 0.8$ which is not negligible on the scale from 0 to 10, but still less than the within-individual standard deviation of $\sqrt{1.372} = 1.17$.

The effect of birth year is significant, with a t -ratio of $-0.394/0.078 = 5.1$, $p < 0.0001$. Here also the effect size is rather small, however. The difference between lowest and highest value of Z (birth year divided by 10) is $1.1 - (-1.1) = 2.2$. Given the presence of an interaction, the main effect α of birth year corresponds to the value $t = t_0$ where the interaction parameter γ in (15.9) cancels, that is, 55 years of age. Thus the main effect of birth year translates to a difference in expected life satisfaction, at 55 years of age, equal to $2.2 \times 0.394 = 0.87$.

The birth year \times age interaction is significant, $t = 0.049/0.019 = 2.6$, $p < 0.01$. The contribution of 0.049 is of the same order of magnitude as the rather irregular differences between the estimated means μ_1 to μ_6 . Those born later on average experience slightly higher life satisfaction during the period when they are 55–60 years old, those born earlier slightly lower or about the same.

The fitted covariance for the complete data vector under Model 4 has elements given by (15.8). Inserting the estimated parameters τ_0^2 , τ_1^2 , and τ_{01} from Table 15.2 into (15.8) yields the covariance matrix

$$\hat{\Sigma}(Y^c) = \begin{pmatrix} 3.67 & 2.16 & 2.06 & 1.96 & 1.87 & 1.77 \\ 2.16 & 3.50 & 2.01 & 1.94 & 1.87 & 1.80 \\ 2.06 & 2.01 & 3.38 & 1.92 & 1.87 & 1.82 \\ 1.96 & 1.94 & 1.92 & 3.31 & 1.87 & 1.85 \\ 1.87 & 1.87 & 1.87 & 1.87 & 3.29 & 1.88 \\ 1.77 & 1.80 & 1.82 & 1.85 & 1.88 & 3.33 \end{pmatrix} \quad (15.10)$$

and the correlation matrix

$$\hat{R}(Y^c) = \begin{pmatrix} 1.00 & 0.60 & 0.58 & 0.56 & 0.54 & 0.51 \\ 0.60 & 1.00 & 0.58 & 0.57 & 0.55 & 0.53 \\ 0.58 & 0.58 & 1.00 & 0.57 & 0.56 & 0.54 \\ 0.56 & 0.57 & 0.57 & 1.00 & 0.57 & 0.56 \\ 0.54 & 0.55 & 0.56 & 0.57 & 1.00 & 0.57 \\ 0.51 & 0.53 & 0.54 & 0.56 & 0.57 & 1.00 \end{pmatrix}$$

These matrices show that the variance does not change a lot, and correlations attenuate slightly as age differences increase, but they are close to the intra-subject correlation estimated above under the compound symmetry model as $\hat{\rho}_I = 0.58$.

However, these values are conditional on the validity of the model with one random slope. We return to these data below, and will investigate the adequacy of this model by testing it against the fully multivariate model.

15.1.3 The fully multivariate model

What is the use of restrictions, such as compound symmetry, on the covariance matrix of a vector of longitudinal measurements? There was a time (say, before 1980) when the

compound symmetry model was used for repeated measures because, in practice, it was impossible to get results for other models. At that time, a complete data matrix was also required. These limitations were gradually overcome between 1970 and 1990.

A more compelling argument for restrictions on the covariance matrix is that when the amount of data is limited, the number of statistical parameters should be kept small to prevent overfitting and to avoid convergence problems in the calculation of the estimates. Another, more appealing, argument is that sometimes the parameters of the models with restricted covariance matrices have nice interpretations. This is the case, for example, for the random slope variance of model (15.7). But when there are enough data, one can also fit a model without restrictions on the covariance matrix, the *fully multivariate model*. This also provides a benchmark to assess the goodness of fit of the models that do have restrictions on the covariance matrix.

Some insight in the extent to which a given model constrains the covariance matrix is obtained by looking at the number of parameters. The covariance matrix of an m -dimensional vector has $m(m + 1)/2$ free parameters (m variances and $m(m - 1)/2$ covariances). The covariance matrix for the compound symmetry model has only 2 parameters. The model with one random slope has 4 parameters; the model with q random slopes has $\{(q + 1)(q + 2)/2\} + 1$ parameters, namely, $(q + 1)(q + 2)/2$ parameters for the random part at level two and 1 parameter for the variance of the random residual. This shows that using some random slopes will quickly increase the number of parameters and thus lead to a better-fitting covariance matrix. The maximum number of random slopes in a conventional random slope model for the fixed occasions design is $q = m - 2$, because with $q = m - 1$ there is one parameter too many.

This suggests how to achieve a perfect fit for the covariance matrix. The fully multivariate model is formulated as a model with a random intercept and $m - 1$ random slopes at level two, and without a random part at level one. Alternatively, the random part at level two may consist of m random slopes and no random intercept. When all variances and covariances between m random slopes are free parameters, the number of parameters in the random part of the model is indeed $m(m + 1)/2$.

The fully multivariate model is little more than a tautology,

$$Y_{ti} = \text{fixed part} + U_{ti}. \quad (15.11)$$

This model is reformulated more recognizably as a hierarchical linear model by the use of dummy variables indicating the measurement occasions. These dummies d_{hti} were defined in (15.2). This leads to the formulation

$$Y_{ti} = \text{fixed part} + \sum_{h=1}^m U_{hi} d_{hti}. \quad (15.12)$$

The variables U_{ti} for $t = 1, \dots, m$ are random at level two, with expectations 0 and an unconstrained covariance matrix. (This means that all variances and all covariances must be freely estimated from the data.) This model does not have a random part at level one. It follows immediately from (15.11) that the covariance matrix of the complete data vector, conditional on the explanatory variables, is identical to the covariance matrix of (U_{1i}, \dots, U_{pi}) . This model for the random part is *saturated* in the sense that it yields a perfect fit for the covariance matrix.

The multivariate model of this section can also be applied to data that do not have a longitudinal nature. This means that the m variables can be completely different, measured on different scales, provided that they have (approximately) a multivariate normal distribution. Thus, this model and the corresponding multilevel software enable the multivariate analysis of incomplete data (provided that missingness is at random as defined in Chapter 9).

Example 15.4 Incomplete paired data.

For the comparison of the means of two variables measured for the same individuals, the paired-samples t -test is a standard procedure. The fully multivariate model provides the possibility of carrying out a similar test if the data are incomplete. This is an example of the simplest multilevel data structure: all level-two units (individuals) contain either one or two level-one units (measurements). Such a data structure may be called an incomplete pretest–post-test design if one measurement was taken before, and the other after, some kind of intervention.

With a different sample from the SOEP, we now consider the question whether average life satisfaction at age 60 differs from the value at age 50. We specify this for persons born in the years 1930–1935, and alive in 1984, and consider their responses at ages 50 and 60. Of the 591 respondents for whom a measurement for at least one of these ages is available, there are 136 who have measurements for both ages, 112 who have a measurement only for age 50 ($t = 0$), and 343 who have a measurement only for age 60 ($t = 1$). With the multilevel approach to multivariate analysis, all these data can be used. The REML estimation method is used (cf. Section 4.7) because this will reproduce exactly the conventional t -test if the data are complete (i.e., all individuals have measurements for both variables).

The model fitted is

$$Y_{ti} = \gamma_0 + \gamma_1 d_{1ti} + U_{ti},$$

where the dummy variable d_1 equals 1 or 0, respectively, depending on whether or not $t = 1$. The null hypothesis that the means are identical for the two time points can be represented by ‘ $\gamma_1 = 0$ ’.

Table 15.3: Estimates for incomplete paired data.

Fixed effect	Coefficient	S.E.
γ_0 Constant term	7.132	0.125
γ_1 Effect time 1	-0.053	0.141
Deviance	2,953	

The results are in Table 15.3. The estimated covariance matrix of the complete data vector is

$$\widehat{\Sigma}(Y^c) = \begin{pmatrix} 4.039 & 1.062 \\ 1.062 & 3.210 \end{pmatrix}.$$

The test of the equality of the two means, which is the test of γ_1 , is not significant ($t = -0.053/0.141 = 0.38$).

Example 15.5 Fully multivariate model for life satisfaction.

Continuing the example of life satisfaction, we now fit the fully multivariate model to the data for the six time points $t = 1, \dots, 6$ corresponding to ages 55, ..., 60. First we consider the model that may be called the *two-level multivariate empty model*, of which the fixed part contains only the dummies

Table 15.4: The empty multivariate model for life satisfaction.

Fixed effect	Coefficient	S.E.
μ_1 Mean at age 55	6.883	0.055
μ_2 Mean at age 56	6.954	0.056
μ_3 Mean at age 57	7.025	0.054
μ_4 Mean at age 58	6.913	0.058
μ_5 Mean at age 59	6.906	0.058
μ_6 Mean at age 60	7.011	0.056
Deviance	21,702.42	

for the effects of the measurement occasions. This is an extension of Model 3 of Table 15.2. The estimates of the fixed effects are presented in Table 15.4.

The estimated covariance matrix of the complete data vector is

$$\widehat{\Sigma}(Y^c) = \begin{pmatrix} 3.68 & 2.17 & 1.97 & 1.95 & 1.88 & 1.65 \\ 2.17 & 3.64 & 2.19 & 1.97 & 1.92 & 1.84 \\ 1.97 & 2.19 & 3.28 & 1.98 & 1.98 & 1.78 \\ 1.95 & 1.97 & 1.98 & 3.50 & 2.08 & 1.87 \\ 1.88 & 1.92 & 1.98 & 2.08 & 3.31 & 1.88 \\ 1.65 & 1.84 & 1.78 & 1.87 & 1.88 & 2.91 \end{pmatrix} \quad (15.13)$$

These estimates are the ML estimates (cf. Section 4.7) of the parameters of a multivariate normal distribution with incomplete data. They differ slightly from the estimates that would be obtained by computing means and variances from available data with pairwise deletion of missing values. The ML estimates are more efficient in the sense of having smaller standard errors.

An eyeball comparison with the fitted covariance matrix for the model with one random slope, given in (15.10), shows that the differences are minor. The variances (diagonal elements) in (15.10) are a smoother function of age than in (15.13), because the model with one random slope implies that the variance must be a quadratic variable on age; see (15.8).

The deviance difference is $\chi^2 = 21,748.03 - 21,702.42 = 45.61$, $df = 11$ (the covariance matrix of Model 3 has four free parameters, this covariance matrix has 15), with $p < 0.0001$. Thus, the fully multivariate model results in a fit that is significantly better, but the estimated covariance matrix is not strongly different. Again, this is a consequence of the large sample size. An advantage of the random slope model is the clearer interpretation of the random part in terms of between-subject differences, as was discussed in Example 15.3. Such an interpretation is not directly obvious from the results of the fully multivariate model.

As a next step the effect of employment situation on life satisfaction is studied, while also taking into account the birth year included in Model 4 of Table 15.2. Employment situation is a changing explanatory variable, measured at the same moments as the dependent variable. There are three categories: working full-time (reference category), working part-time, and not working. As this is a level-one variable, the issue of the difference between within-group and between-group regressions, treated in Section 4.6, comes up again. In the longitudinal case, the analog of the group mean is the person mean. Including this in the model would, however, imply an effect from events that occur in the future because, at each moment except the last, the person mean of a changing explanatory variable depends on measurements to be made at a future point in time. Therefore we do not include

the person mean but, rather, the employment situation at the start of the study period, that is, at 55 years of age. This is regarded as a baseline situation, and the effects of current employment situation are interpreted as effects of changes with respect to this baseline.

Table 15.5: The multivariate model with the effects of cohort and employment situation.

Fixed effect	Coefficient	S.E.
μ_1 Effect of age 55	7.064	0.059
μ_2 Effect of age 56	7.136	0.061
μ_3 Effect of age 57	7.215	0.060
μ_4 Effect of age 58	7.117	0.064
μ_5 Effect of age 59	7.129	0.066
μ_6 Effect of age 60	7.257	0.068
α_1 Main effect birth year	-0.369	0.075
α_2 Interaction birth year \times age	0.042	0.019
α_3 Working part-time at age 55	-0.247	0.127
α_4 Not working at age 55	-1.171	0.142
α_5 Currently working part-time	-0.199	0.064
α_6 Currently not working	-0.112	0.076
Deviance	21,571.97	

The estimates of the fixed effects are presented in Table 15.5. The effect of birth year and its interaction with age is similar to Model 4 of Table 15.2. As to employment situation, compared to full-time working, not working at age 55 has a strong negative effect on life satisfaction ($\hat{\alpha}_4 = -1.171$, $t = -1.171/0.142 = -8.24$, $p < 0.0001$), and working part-time at age 55 a weak negative effect ($\hat{\alpha}_3 = -0.247$, $t = -0.247/0.127 = -1.95$, $p \approx 0.05$). Currently working part-time (compared to full-time) also has a weak negative effect.

Concluding remarks on the fully multivariate model

There is nothing special about this formulation as a hierarchical linear model of a fully multivariate model for repeated measures with a fixed occasion design. It is just a mathematical formula. What is special is that available algorithms and software for multilevel analysis accept this formulation, even without a random part at level one, and can calculate ML or REML parameter estimates. In this way, multilevel software can compute ML or REML estimates for multivariate normal distributions with incomplete data, and also for multivariate regression models with incomplete data and with sets of explanatory variables that are different for the different dependent variables. Methods for such models have been in existence for a longer period (see Little and Rubin, 2002), but their practical use has been much facilitated by the development of multilevel software.

The use of the dummy variables (15.2) has the nice feature that the covariance matrix obtained for the random slopes is exactly the covariance matrix for the complete data vector. However, one could also give the random slopes to other variables. The only requirement is that there are random slopes for m linearly independent variables, depending only on the

measurement occasion and not on the individual. For example, one could use powers of $(t - t_0)$, where t_0 may be any meaningful reference point, such as the average of all values for t . This means that one uses variables

$$\begin{aligned} d_{1ti} &= 1, \\ d_{hti} &= (t - t_0)^{h-1}, \quad h = 2, \dots, m, \end{aligned}$$

still with model specification (15.12). Since the first ‘variable’ is constant, this effectively means that one uses a random intercept and $m - 1$ random slopes.

Each model for the random part with a restricted covariance matrix is a submodel of the fully multivariate model. Therefore the fit of such a restricted model can be tested by comparing it to the fully multivariate model by means of a likelihood ratio (deviance) test (see Chapter 6).

For complete data, an alternative to the hierarchical linear model exists in the form of multivariate analysis of variance (MANOVA) and multivariate regression analysis. This is documented in many textbooks (e.g., Maxwell and Delaney, 2004; Stevens, 2009) and implemented in standard software such as SPSS and SAS. The advantage of these methods is the fact that under the assumption of multivariate normal distributions the tests are exact, whereas the tests in the hierarchical linear model formulation are approximate. For incomplete multivariate data, however, exact methods are not available. Maas and Snijders (2003) elaborate the correspondence between the MANOVA approach and the hierarchical linear model approach.

15.1.4 Multivariate regression analysis

Because of the unrestricted multivariate nature of the fully multivariate model, it is not required that the outcome variables Y_{ti} are repeated measurements of conceptually the same variable. This model is applicable to any set of multivariate measurements for which a multivariate normal distribution is an adequate model. Thus, the multilevel approach yields estimates and tests for normally distributed multivariate data with randomly missing observations (see also Maas and Snijders, 2003).

The fully multivariate model is also the basis for multivariate regression analysis, but then the focus usually is on the fixed part. There are supposed to be individual-dependent explanatory variables Z_1, \dots, Z_q . If the regression coefficients of the m dependent variables on the Z_h are all allowed to be different, the regression coefficient of outcome variable Y_t on explanatory variable Z_h being denoted by γ_{ht} , then the multivariate regression model with possibly incomplete data can be formulated as

$$Y_{ti} = \mu_t + \sum_{l=1}^m \sum_{h=1}^q \gamma_{ht} d_{hti} z_{hi} + \sum_{h=1}^m U_{hi} d_{hti}. \quad (15.14)$$

This shows that the occasion dummies are fundamental, not only because they have random slopes, but also because in the fixed part all variables are multiplied by these dummies. If some of the Z_k variables are not to be used for all dependent variables, then the corresponding cross-product terms $d_{hti}z_{hi}$ can be dropped from (15.14).

15.1.5 Explained variance

The proportion of explained variance at level one can be defined, analogous to explained variance for grouped data treated in Section 7.1, as the proportional reduction in prediction error for individual measurements, averaged over the m measurement occasions. A natural baseline model is provided by a model which can be represented by

$$Y_{it} = \mu_t + \text{random part},$$

that is, the fixed part depends on the measurement occasion but not on other explanatory variables, and the random part is chosen so as to provide a good fit to the data.

The proportion of explained variance at level one, R_1^2 , is then the proportional reduction in the average residual variance,

$$\frac{1}{m} \sum_{t=1}^m \text{var}(Y_{ti}),$$

when going from the baseline model to the model containing the explanatory variables.

If the random part of the compound symmetry model is adequate, the baseline model is (15.1). In this case the definition of R_1^2 is just as in Section 7.1.

If the compound symmetry model is not adequate one could use, at the other end of the spectrum, the multivariate random part. This yields the baseline model

$$Y_{ti} = \mu_t + U_{ti} = \sum_{h=1}^m \mu_h d_{hti} + \sum_{h=1}^m U_{hi} d_{hti}$$

which is the fully multivariate model without covariates; cf. (15.11) and (15.12).

In all models for fixed occasion designs, the calculation of the proportions of explained variance can be related to the fitted complete data covariance matrix, $\widehat{\Sigma}(Y^c)$. The value of R_1^2 is the proportional reduction in the sum of diagonal values of this matrix when going from the baseline model to the model including the explanatory variables.

Example 15.6 Explained variance for life satisfaction.

We continue the example of life satisfaction of people aged 55–60, now computing the proportion of variance explained by cohort (birth year), employment situation, and the interaction of the cohort with age as included in the model of Table 15.5; to which are added the total income after taxes, logarithmically transformed, and satisfaction with health, also measured on a scale from 0 to 10. The last two covariates were centered to have overall means 0. Their standard deviations are 0.67 and 2.3, respectively.

First suppose that the compound symmetry model is used. The baseline model then is Model 2 in Table 15.1. The variance per measurement is $1.991 + 1.452 = 3.443$. The results of the compound symmetry model with the effect of cohort and employment situation are given in Table 15.6. In comparison with Model 2 of Table 15.1, inclusion of these fixed effects has led especially to a smaller variance at the individual level; these variables, although some of them are changing covariates, contribute little to explaining year-to-year fluctuations in life satisfaction. The residual variance per measurement is $1.038 + 1.351 = 2.389$. Thus, the proportion of variance explained at level one is $R_1^2 = 1 - (2.389/3.443) = 0.306$.

Table 15.6: The compound symmetry model with the effects of cohort, employment situation, income, and health satisfaction.

Fixed effect	Coefficient	S.E.
μ_1 Effect of age 55	6.925	0.050
μ_2 Effect of age 56	7.017	0.052
μ_3 Effect of age 57	7.113	0.053
μ_4 Effect of age 58	6.999	0.056
μ_5 Effect of age 59	6.997	0.059
μ_6 Effect of age 60	7.121	0.065
α_1 Main effect birth year	-0.426	0.062
α_2 Interaction birth year \times age	0.028	0.017
α_3 Working part-time at age 55	-0.138	0.104
α_4 Not working at age 55	-0.521	0.120
α_5 Currently working part-time	-0.090	0.061
α_6 Currently not working	-0.067	0.072
α_7 In(Income after taxes)	0.167	0.041
α_8 Satisfaction with health	0.301	0.010
Random effect	Parameter	S.E.
<i>Level-two (i.e., individual) variance:</i>		
$\tau_0^2 = \text{var}(U_{0i})$	1.038	0.055
<i>Level-one (i.e., occasion) variance:</i>		
$\sigma^2 = \text{var}(R_{ti})$	1.351	0.028
Deviance	21,648.76	

Next suppose that the fully multivariate model is used. The estimated fixed effects do not change much. The estimated covariance matrix in the fully multivariate model is given in (15.13), while the residual covariance matrix of the model with these fixed effects, as estimated by employing the fully multivariate model, is

$$\widehat{\Sigma}(Y^c) = \begin{pmatrix} 2.51 & 1.15 & 1.04 & 1.01 & 0.95 & 0.75 \\ 1.15 & 2.66 & 1.21 & 1.03 & 0.98 & 1.03 \\ 1.04 & 1.21 & 2.27 & 1.08 & 1.08 & 0.90 \\ 1.01 & 1.03 & 1.08 & 2.45 & 1.07 & 0.95 \\ 0.95 & 0.98 & 1.08 & 1.07 & 2.23 & 0.90 \\ 0.75 & 1.03 & 0.90 & 0.95 & 0.90 & 2.01 \end{pmatrix}. \quad (15.15)$$

The sum of diagonal values is 20.32 for the first covariance matrix and 14.13 for the second. Hence, calculated on the basis of the fully multivariate model, $R_1^2 = 1 - (14.13/20.32) = 0.305$. Comparing this to the values obtained above shows that for the calculation of R_1^2 it does not make much difference which random part is used. The calculations using the fully multivariate model are more reliable, but in most cases the simpler calculations using the compound symmetry ('random intercept') model will lead to almost the same values for R_1^2 .

15.2 Variable occasion designs

In data collection designs with variable measurement occasions, there is no such thing as a complete data vector. The data are ordered according to some underlying dimension (e.g. time or age), and, for each individual, data are recorded at some set of time points not necessarily related to the time points at which the other individuals are observed. For example, body heights are recorded at a number of moments during childhood and adolescence, the moments being determined by convenience rather than strict planning.

The same notation can be used as in the preceding section: for individual i , the dependent variable Y_{ti} is measured at occasions $s = 1, \dots, m_i$. The time of measurement for Y_{ti} is t . This ‘time’ variable can refer to age, clock time, etc., but also to some other dimension such as one or more spatial dimensions, the concentration of a poison, etc.

The number of measurements per individual, m_i , can be anything. It is not a problem that some individuals contribute only one observation. This number must not in itself be informative about the process studied, however. Therefore it is not allowed that the observation times t are defined as the moments when some event occurs (e.g., change of job); such data collection designs should be investigated using event history models (e.g., Singer and Willett, 2003, Part II; Mills, 2011). Larger numbers m_i give more information about intra-individual differences, of course, and with larger average m_i s one will be able to fit models with a more complicated, and more precise, random part.

Because of the unbalanced nature of the data set, the random intercept and slope models are easily applicable, but the other models of the preceding section either have no direct analog, or an analog that is considerably more complicated. This section is restricted to random slope models, and follows the approach of Snijders (1996), where further elaboration and background material may be found. Other introductions to multilevel models for longitudinal data with variable occasion designs may be found, for example, in Raudenbush (1995) and Singer and Willett (2003, Chapter 5).

15.2.1 Populations of curves

An attractive way to view repeated measures in a variable occasion design, ordered according to an underlying dimension t (referred to as time), is as observations on a *population of curves*. (This approach can be extended to two- or higher-dimensional ordering principles.) The variable Y for individual i follows a development represented by the function $F_i(t)$, and the population of interest is the population of curves F_i . The observations yield snapshots of this function on a finite set of time points, with superimposed residuals R_{ti} which represent incidental deviations and measurement error:

$$Y_{ti} = F_i(t) + R_{ti}, \quad s = 1, \dots, m_i. \quad (15.16)$$

Statistical modeling consists of determining an adequate class of functions F_i and investigating how these functions depend on explanatory variables.

15.2.2 Random functions

It makes sense here to consider models with functions of t as the only explanatory variables. This can be regarded as a model for random functions that has the role of a baseline model comparable to the role of the empty model in Chapter 4. Modeling can proceed by

first determining an adequate random function model and subsequently incorporating the individual-based explanatory variables.

One could start by fitting the empty model, that is, the random intercept model with no explanatory variables, but this is such a trivial model when one is modeling curves that it may also be skipped. The next simplest model is a linear function,

$$F_i(t) = \beta_{0i} + \beta_{1i}(t - t_0).$$

The value t_0 is subtracted for the same reasons as in model (15.7) for the fixed occasion design. It may be some reference value within the range of observed values for t .

If the β_{hi} ($h = 0, 1$) are split into their population average¹ γ_{h0} and the individual deviation $U_{hi} = \beta_{hi} - \gamma_{h0}$, similar to (5.2), this gives the following model for the observations:

$$Y_{ti} = \gamma_{00} + \gamma_{10}(t - t_0) + U_{0i} + U_{1i}(t - t_0) + R_{ti}. \quad (15.17)$$

The population of curves is now characterized by the bivariate distribution of the intercepts β_{0i} at time $t = t_0$, and the slopes β_{1i} . The average intercept at $t = t_0$ is γ_{00} and the average slope is γ_{10} . The intercepts have variance $\tau_0^2 = \text{var}(U_{0i})$, the slopes have variance $\tau_1^2 = \text{var}(U_{1i})$, and the intercept-slope covariance is $\tau_{01} = \text{cov}(U_{0i}, U_{1i})$. In addition, the measurements exhibit random deviations from the curve with variance $\sigma^2 = \text{var}(R_{ti})$. This level-one variance represents deviations from linearity together with measurement inaccuracy, and can be used to assess the fit of linear functions to the data.

Even when this model fits only moderately well, it can have a meaningful interpretation, because the estimated slope and slope variance still will give an impression of the average increase of Y per unit of time.

Example 15.7 Retarded growth.

In several examples in this section we present results for a data set of children with retarded growth. More information about this research can be found in Rekers-Mombarg et al. (1997). The data set concerns children who saw a pediatric endocrinologist because of growth retardation. Height measurements are available at irregular ages varying between 0 and 30 years. In the present example we consider the measurements for ages from 5.0 to 10.0 years. The linear growth model represented by (15.17) is fitted to the data.

For this age period there are data for 336 children available, establishing a total of 1,886 height measurements. Height is measured in centimeters (cm). Age is measured in years and the reference age is chosen as $t_0 = 5$ years, so the intercept refers to heights at age 5. Parameter estimates for model (15.17) are presented in Table 15.7.

The level-one standard deviation is so low ($\hat{\sigma} = 0.9$ cm) that it can be concluded that the fit of the linear growth model for the period from 5 to 10 years is quite adequate. Deviations from linearity will be usually smaller than 2 cm (which is slightly more than two standard errors). This is notwithstanding the fact that, given the large sample size, it is possible that the fit can be improved significantly from a statistical point of view by including nonlinear growth terms.

The intercept parameters show that at 5 years, these children have an average height of 96.3 cm with a variance of $\hat{\tau}_0^2 + \hat{\sigma}^2 = 19.79 + 0.82 = 20.61$ and an associated standard deviation of $\sqrt{20.61} = 4.5$ cm. The growth per year is $\gamma_{10} + U_{1i}$, which has an estimated mean of 5.5 cm and

¹The notation with the parameters γ has nothing to do with the γ parameters used earlier in this chapter in (15.4), but is consistent with the notation in Chapter 5.

Table 15.7: Linear growth model for 5–10-year-old children with retarded growth.

Fixed effect	Coefficient	S.E.
γ_{00} Intercept	96.32	0.285
γ_{10} Age	5.53	0.08
Random effect	Parameter	S.E.
<i>Level-two (i.e., individual) random effects:</i>		
τ_0^2 Intercept variance	19.79	1.91
τ_1^2 Slope variance for age	1.65	0.16
τ_{01} Intercept–slope covariance	-3.26	0.46
<i>Level-one (i.e., occasion) variance:</i>		
σ^2 Residual variance	0.82	0.03
Deviance	7,099.87	

standard deviation of $\sqrt{1.65} = 1.3$ cm. This implies that 95% of these children have growth rates, averaged over this five-year period, between $5.5 - 2 \times 1.3 = 2.9$ and $5.5 + 2 \times 1.3 = 8.1$ cm per year. The slope–intercept covariance is negative: children who are relatively short at 5 years grow relatively fast between 5 and 10 years.

Polynomial functions

To obtain a good fit, however, one can try and fit more complicated random parts. One possibility is to use a *polynomial random part*. This means that one or more powers of $(t - t_0)$ are given a random slope. This corresponds to polynomials for the function F_i ,

$$F_i(t) = \beta_{0i} + \beta_{1i}(t - t_0) + \beta_{2i}(t - t_0)^2 + \dots + \beta_{ri}(t - t_0)^r, \quad (15.18)$$

where r is a suitable number, called the degree of the polynomial. For example, one may test the null hypothesis that individual curves are linear by testing the null hypothesis $r = 1$ against the alternative hypothesis that the curves are quadratic, $r = 2$. Any function for which the value is determined at m points can be represented exactly by a polynomial of degree $m - 1$. Therefore, in the fixed occasion model with m measurement occasions, it makes no sense to consider polynomials of degree higher than $m - 1$.

Usually the data contain more information about inter-individual differences (corresponding to the fixed part) than about intra-individual differences (the random part). For some of the coefficients β_{hi} in (15.18), there will be empirical evidence that they are nonzero, but not that they are variable across individuals. Therefore, as in Section 5.2.2, one can have a fixed part that is more complicated than the random part and give random slopes only to the lower powers of $(t - t_0)$. This yields the model

$$Y_{ti} = \gamma_{00} + \sum_{h=1}^r \gamma_{h0}(t - t_0)^h + U_{0i} + \sum_{h=1}^p U_{hi}(t - t_0)^h + R_{ti}, \quad (15.19)$$

which has the same structure as (5.15). For $h = p + 1, \dots, r$, parameter γ_{h0} is the value of the coefficient β_{hi} , constant over all individuals i . For $h = 1, \dots, p$, the coefficients β_{hi} are individual-dependent, with population average γ_{h0} and individual deviations $U_{hi} = \beta_{hi} - \gamma_{h0}$. All variances and covariances of the random effects U_{hi} are estimated from the data. The mean curve for the population is given by

$$\mathcal{E}(F_i(t)) = \gamma_{00} + \sum_{h=1}^r \gamma_{h0}(t - t_0)^h. \quad (15.20)$$

Numerical difficulties appear less often in the estimation of models of this kind when t_0 has a value in the middle of the range of t -values in the data set than when t_0 is outside or at one of the extremes of this range. Therefore, when convergence problems occur, it is advisable to try and work with a value of t_0 close to the average or median value of t . Changing the value of t_0 only amounts to a new parametrization, that is, a different t_0 leads to different parameters γ_{h0} for which, however, formula (15.20) constitutes the same function and the deviance of the model (given that the software will calculate it) is also the same.

The number of random slopes, p , is not greater than r and may be considerably smaller. To give a rough indication, $r + 1$ may not be larger than the total number of distinct time points, that is, the number of different values of t in the entire data set for which observations exist; also it should not be larger than a small proportion, say, 10%, of the total number of observations $\sum_i m_i$. On the other hand, p will rarely be much larger than the maximum number of observations per individual, $\max_i m_i$.²

Example 15.8 Polynomial growth model for children with retarded growth.

We continue the preceding example of children with retarded growth for which height measurements are considered in the period between the ages of 5 and 10 years. In fitting polynomial models, convergence problems occurred when the reference age t_0 was chosen as 5 years, but not when it was chosen as the midpoint of the range, 7.5 years.

A cubic model, that is, a model with polynomials of the third degree, turned out to yield a much better statistical fit than the linear model of Table 15.7. Parameter estimates are shown in Table 15.8 for the cubic model with $t_0 = 7.5$ years. So the intercept parameters refer to children of this age.

For the level-two random slopes ($U_{0i}, U_{1i}, U_{2i}, U_{3i}$) the estimated correlation matrix is

$$\widehat{R}_U = \begin{pmatrix} 1.0 & 0.17 & -0.27 & 0.04 \\ 0.17 & 1.0 & 0.11 & -0.84 \\ -0.27 & 0.11 & 1.0 & -0.38 \\ 0.04 & -0.84 & -0.38 & 1.0 \end{pmatrix}.$$

The fit is much better than that of the linear model (deviance difference 496.12 for 9 degrees of freedom). The random effect of the cubic term is significant (the model with fixed effects up to the power $r = 3$ and with $p = 2$ random slopes, not shown in the table, has deviance 6,824.63, so the deviance difference for the random slope of $(t - t_0)^3$ is 221.12 with 4 degrees of freedom).

²In the fixed occasion design, the number of random effects, including the random intercept, cannot be greater than the number of measurement occasions. In the variable occasion design this strict upper bound does not figure, because the variability of time points of observations can lead to richer information about intra-individual variations. But if one obtains a model with clearly more random slopes than the maximum of all m_i , this may mean that one has made an unfortunate choice of the functions of time (polynomial or other) that constitute the random part at level two, and it may be advisable to try and find another, smaller, set of functions of t for the random part with an equally good fit to the data.

Table 15.8: Cubic growth model for 5–10-year-old children with retarded growth.

Fixed effect	Coefficient	S.E.
γ_{00} Intercept	110.40	0.22
γ_{10} $t - 7.5$	5.23	0.12
γ_{20} $(t - 7.5)^2$	-0.007	0.038
γ_{30} $(t - 7.5)^3$	0.009	0.020
Random effect	Variance	S.E.
<i>Level-two (i.e., individual) random effects:</i>		
τ_0^2 Intercept variance	13.80	1.19
τ_1^2 Slope variance $t - t_0$	2.97	0.32
τ_2^2 Slope variance $(t - t_0)^2$	0.255	0.032
τ_3^2 Slope variance $(t - t_0)^3$	0.066	0.009
<i>Level-one (i.e., occasion) variance:</i>		
σ^2 Residual variance	0.37	0.02
Deviance	6,603.75	

The mean curve for the population (cf. (15.20)) is given here by

$$\mathcal{E}(F_i(t)) = 110.40 + 5.23(t - 7.5) - 0.007(t - 7.5)^2 + 0.009(t - 7.5)^3.$$

This deviates hardly at all, and not significantly, from a straight line. However, the individual growth curves do differ from straight lines, but the pattern of variation implied by the level-two covariance matrix is quite complex. We return to this data set in the next example.

Other functions

There is nothing sacred about polynomial functions. They are convenient, reasonably flexible, and any reasonably smooth function can be approximated by a polynomial if only you are prepared to use a polynomial of sufficiently high degree. One argument for using other classes of functions is that some function shapes are approximated more parsimoniously by other functions than polynomials. Another argument is that polynomials are wobbly: when the value of a polynomial function $F(t)$ is changed a bit at one value of t , this may require coefficient changes that make the function change a lot at other values of t . In other words, the fitted value at any given value of t can depend strongly on observations for quite distant values of t . This kind of sensitivity is often undesirable.

One can use other functions instead of polynomials. If the functions used are called $f_1(t), \dots, f_p(t)$, then instead of (15.19) the random function is modeled as

$$F_i(t) = \beta_{0i} + \sum_{h=1}^r \beta_{hi} f_h(t) \tag{15.21}$$

and the observations as

$$Y_{ti} = \gamma_{00} + \sum_{h=1}^r \gamma_{h0} f_h(t) + U_{0i} + \sum_{h=1}^p U_{hi} f_h(t) + R_{ti}. \quad (15.22)$$

What is a suitable class of functions depends on the phenomenon under study and the data at hand. Random functions that can be represented by (15.22) are particularly convenient because this representation is a *linear* function of the parameters γ and the random effects U . Therefore, (15.22) defines an instance of the hierarchical linear model. Below we treat various classes of functions; the choice among them can be based on the deviance but also on the level-one residual variance σ_R^2 , which indicates the size of the deviations of individual data points with respect to the fitted population of functions.

Sometimes, however, theory or data point to function classes where the statistical parameters enter in a nonlinear fashion. In this chapter we restrict attention to linear models, that is, models that can be represented as (15.22). Readers interested in nonlinear models are referred to more specialized literature such as Davidian and Giltinan (1995), Hand and Crowder (1996, Chapter 8), and Singer and Willett (2003, Chapter 6).

Piecewise linear functions

A class of functions which is flexible, easy to comprehend, and for which the fitted function values have a very restricted sensitivity for observations made at other values of t , is the class of continuous *piecewise linear functions*. These are continuous functions whose slopes may change discontinuously at a number of values of t called *nodes*, but which are linear (and hence have constant slopes) between these nodes. A disadvantage is their angular appearance.

The basic piecewise linear function is linear on a given interval (t_1, t_2) and constant outside this interval, as defined by

$$f(t) = \begin{cases} a & (t \leq t_1) \\ a + (b - a) \frac{t - t_1}{t_2 - t_1} & (t_1 < t < t_2) \\ b & (t \geq t_2). \end{cases} \quad (15.23)$$

Often one of the constant values a and b is chosen to be 0. The nodes are t_1 and t_2 . Boundary cases are the functions (choosing $a = t_1$ and $b = 0$ and letting the lower node t_1 tend to minus infinity)

$$f(t) = \begin{cases} t - t_2 & (t < t_2) \\ 0 & (t \geq t_2) \end{cases}$$

and (choosing $a = 0$ and $b = t_2$ and letting the upper node t_2 tend to plus infinity)

$$f(t) = \begin{cases} 0 & (t \leq t_1) \\ t - t_1 & (t > t_1), \end{cases}$$

and also linear functions such as $f(t) = t$ (where both nodes are infinite). Each piecewise linear function can be obtained as a linear combination of these basic functions. The choice

of nodes sometimes will be suggested by the problem at hand, and in other situations has to be determined by trial and error.

Example 15.9 Piecewise linear models for retarded growth.

Let us try to improve on the polynomial models for the retarded growth data of the previous example by using piecewise linear functions. Recall that the height measurements were considered for ages from 5.0 to 10.0 years. For ease of interpretation, the nodes are chosen at the children's birthdays, at 6.0, 7.0, 8.0, and 9.0 years. This means that growth is assumed to proceed linearly during each year, but that growth rates may be different between the years. For the comparability with the polynomial model, the intercept again corresponds to height at the age of 7.5 years. This is achieved by using piecewise linear functions that all are equal to 0 for $t = 7.5$. Accordingly, the model is based on the following five basic piecewise linear functions:

$$\begin{aligned} f_1(t) &= \begin{cases} -1 & (t \leq 5) \\ t - 6 & (5 < t < 6) \\ 0 & (t \geq 6), \end{cases} \\ f_2(t) &= \begin{cases} -1 & (t \leq 6) \\ t - 7 & (6 < t < 7) \\ 0 & (t \geq 7), \end{cases} \\ f_3(t) &= \begin{cases} -0.5 & (t \leq 7) \\ t - 7.5 & (7 < t < 8) \\ 0.5 & (t \geq 8), \end{cases} \\ f_4(t) &= \begin{cases} 0 & (t \leq 8) \\ t - 8 & (8 < t < 9) \\ 1 & (t \geq 9), \end{cases} \\ f_5(t) &= \begin{cases} 0 & (t \leq 9) \\ t - 9 & (9 < t < 10) \\ 1 & (t \geq 10). \end{cases} \end{aligned}$$

The results for this model are presented in Table 15.9.

The level-two random effects ($U_{0i}, U_{1i}, \dots, U_{5i}$) have estimated correlation matrix

$$\widehat{R}_U = \begin{pmatrix} 1.0 & 0.22 & 0.31 & 0.14 & -0.05 & 0.09 \\ 0.22 & 1.0 & 0.23 & 0.01 & 0.18 & 0.33 \\ 0.31 & 0.01 & 1.0 & 0.12 & -0.16 & 0.48 \\ 0.14 & 0.01 & 0.12 & 1.0 & 0.47 & -0.23 \\ -0.05 & 0.18 & -0.16 & 0.47 & 1.0 & 0.03 \\ 0.09 & 0.33 & 0.48 & -0.23 & 0.03 & 1.0 \end{pmatrix}$$

The average growth rate is between 5 and 6 cm/year over the whole age range from 5 to 10 years. There is large variability in individual growth rates. All slope variances are between 3 and 4, so the between-children standard deviations in yearly growth rate are almost 2. The correlations between individual growth rates in different years are not very high, ranging from -0.23 (between ages 7–8 and 9–10) to 0.48 (between ages 6–7 and 9–10). This indicates that the growth rate fluctuates rather erratically from year to year around the average of about 5.5 cm/year.

The deviance for this piecewise linear model is 121.88 less than the deviance of the polynomial model of Table 15.8, while it has 13 parameters more. This is a large deviance difference for only 13 degrees of freedom. The residual level-one variance also is smaller. Although the polynomial model is not a submodel of the piecewise linear model, so that we cannot test the former against the latter by the usual deviance test, yet we may conclude that the piecewise model not only is more clearly interpretable but also has a better fit than the polynomial model.

Table 15.9: Piecewise linear growth model for 5–10-year-old children with retarded growth.

Fixed effect	Coefficient	S.E.
γ_{00} Intercept	110.40	0.22
$\gamma_{10} f_1$ (5–6 years)	5.79	0.24
$\gamma_{20} f_2$ (6–7 years)	5.59	0.18
$\gamma_{30} f_3$ (7–8 years)	5.25	0.16
$\gamma_{40} f_4$ (8–9 years)	5.16	0.15
$\gamma_{50} f_5$ (9–10 years)	5.50	0.16
Random effect	Variance	S.E.
<i>Level-two (i.e., individual) random effects:</i>		
τ_0^2 Intercept variance	13.91	1.20
τ_1^2 Slope variance f_1	3.97	0.82
τ_2^2 Slope variance f_2	3.80	0.57
τ_3^2 Slope variance f_3	3.64	0.50
τ_4^2 Slope variance f_4	3.42	0.45
τ_5^2 Slope variance f_5	3.77	0.53
<i>Level-one (i.e., occasion) variance:</i>		
σ^2 Residual variance	0.302	0.015
Deviance	6,481.87	

Spline functions

Another flexible class of functions which is suitable for modeling longitudinal measurements within the framework of the hierarchical linear model is the class of *spline functions*. Spline functions are smooth piecewise polynomials. A number of points called *nodes* are defined on the interval where the spline function is defined; between each pair of adjacent nodes the spline function is a polynomial of degree p , while these polynomials are glued together so smoothly that the function itself and its derivatives, of order up to $p - 1$, are also continuous at the nodes. For $p = 1$ this leads again to piecewise linear functions, but for $p > 1$ it yields functions which are smooth and do not have the kinky appearance of piecewise linear functions. An example of a quadratic spline (i.e., a spline with $p = 2$) was given in Chapter 8 by equation (8.3) and Figure 8.1. This equation and figure represent a function of IQ which is a quadratic for $\text{IQ} < 0$ and also for $\text{IQ} > 0$, but which has different coefficients for these two domains. The point 0 here is the node. The coefficients are such that the function and its derivative are also continuous at the node, as can be seen from the graph. Therefore it is a spline. Cubic splines ($p = 3$) are also often used. We present here only a sketchy introduction to the use of splines. For a more elaborate introduction to the use of spline functions in (single-level) regression models, see Seber and Wild (1989, Section 9.5). The use of spline functions in longitudinal multilevel models was discussed by Pan and Goldstein (1998).

Suppose that one is investigating the development of some characteristic over the age of 12–17 years. Within each of the intervals 12–15 and 15–17 years, the development curves might be approximately quadratic (this could be checked by a polynomial trend analysis for the data for these intervals separately), while they are smooth but not quadratic over the entire range of 12–17. In such a case it would be worthwhile to try a quadratic spline ($p = 2$) with one node, at 15 years. Defining $t_1 = 15$, the basic functions can be taken as

$$\begin{aligned} f_1(t) &= t && \text{(linear function)} \\ f_2(t) &= \begin{cases} (t - t_1)^2 & (t \leq t_1) \\ 0 & (t > t_1) \end{cases} && \text{(quadratic to the left of } t_1\text{)} \\ f_3(t) &= \begin{cases} 0 & (t \leq t_1) \\ (t - t_1)^2 & (t > t_1) \end{cases} && \text{(quadratic to the right of } t_1\text{).} \end{aligned} \quad (15.24)$$

The functions f_2 and f_3 are quadratic functions to the left of t_1 and right of t_1 respectively, and they are continuous and have continuous derivatives. That the functions are continuous and have continuous derivatives even at the node $t = t_1$ can be verified by elementary calculus or by drawing a graph.

The individual development functions are modeled as

$$F_i(t) = \beta_{0i} + \beta_{1i}f_1(t) + \beta_{2i}f_2(t) + \beta_{3i}f_3(t). \quad (15.25)$$

If $\beta_{2i} = \beta_{3i}$, the curve for individual i is exactly quadratic. The freedom to have these two coefficients differ from each other allows us to represent functions that look very different from quadratic functions; for example, if these coefficients have opposite signs then the function will be concave on one side of t_1 and convex on the other side. Equation (8.3) and Figure 8.1 provide an example of exactly such a function, where t is replaced by IQ and the node is the point IQ = 0.

The treatment of this model within the hierarchical linear model approach is completely analogous to the treatment of polynomial models. The functions f_1 , f_2 , and f_3 constitute the fixed part of the model as well as the random part of the model at level two. If there is no evidence for individual differences with respect to the coefficients β_{2i} and/or β_{3i} , then these could be deleted from the random part.

Formula (15.25) shows that a quadratic spline with one node has one parameter more than a quadratic function (4 instead of 3). Each node added further will increase the number of parameters of the function by one. There is considerable freedom of choice in defining the basic functions, subject to the restriction that they are quadratic on each interval between adjacent nodes, and are continuous with continuous derivatives at the nodes. For two nodes, t_1 and t_2 , a possible choice is the following. This representation employs a reference value t_0 that is an arbitrary (convenient or meaningful) value. It is advisable to use a t_0 within the range of observation times in the data. The basic functions are

$$\begin{aligned} f_1(t) &= t - t_0 && \text{(linear function)} \\ f_2(t) &= (t - t_0)^2 && \text{(quadratic function)} \\ f_3(t) &= \begin{cases} (t - t_1)^2 & (t \leq t_1) \\ 0 & (t > t_1) \end{cases} && \text{(quadratic to the left of } t_1\text{)} \\ f_4(t) &= \begin{cases} 0 & (t \leq t_2) \\ (t - t_2)^2 & (t > t_2) \end{cases} && \text{(quadratic to the right of } t_2\text{).} \end{aligned} \quad (15.26)$$

When these four functions are used in the representation

$$F_i(t) = \beta_{0i} + \beta_{1i}f_1(t) + \beta_{2i}f_2(t) + \beta_{3i}f_3(t) + \beta_{4i}f_4(t),$$

coefficient β_{2i} is the quadratic coefficient in the interval between t_1 and t_2 , while β_{3i} and β_{4i} are the changes in the quadratic coefficient that occur when time t passes the nodes t_1 or t_2 , respectively. The quadratic coefficient for $t < t_1$ is $\beta_{2i} + \beta_{3i}$, and for $t > t_2$ it is $\beta_{2i} + \beta_{4i}$.

The simplest cubic spline ($p = 3$) has one node. If the reference point t_0 is equal to the node, then the basic functions are

$$\begin{aligned} f_1(t) &= t - t_0 && \text{(linear function)} \\ f_2(t) &= (t - t_0)^2 && \text{(quadratic function)} \\ f_3(t) &= \begin{cases} (t - t_0)^3 & (t \leq t_0) \\ 0 & (t > t_0) \end{cases} && \text{(cubic to the left of } t_0\text{)} \\ f_4(t) &= \begin{cases} 0 & (t \leq t_0) \\ (t - t_0)^3 & (t > t_0) \end{cases} && \text{(cubic to the right of } t_0\text{).} \end{aligned} \quad (15.27)$$

For more than two nodes, and an arbitrary order p of the polynomials, the basic spline functions may be chosen as follows, for nodes denoted by t_1, \dots, t_M :

$$\begin{aligned} f_k(t) &= (t - t_0)^k && (k = 1, \dots, p) \\ f_{p+k}(t) &= \begin{cases} 0 & (t \leq t_k) \\ (t - t_k)^p & (t > t_k) \end{cases} && (k = 1, \dots, M) \end{aligned} \quad (15.28)$$

The choice of nodes is important to obtain a well-fitting approximation. Since the spline functions are nonlinear functions of the nodes, formal optimization of the node placement is more complicated than fitting spline functions with given nodes, and the interested reader is referred to the literature on spline functions for the further treatment of node placement. If the individuals provide enough observations (i.e., m_i is large enough), plotting observations for individuals, together with some trial and error, can lead to a good placement of the nodes.

Example 15.10 Cubic spline models for retarded growth, 12–17 years.

In this example we again consider the height measurements of children with retarded growth studied by Rekers-Mombarg et al. (1997), but now the focus is on the age range from 12 to 17 years. After deletion of cases with missing data, in this age range there are a total of 1,941 measurements which were taken for a sample of 321 children.

Some preliminary model fits showed that quadratic splines provide a better fit than piecewise linear functions, and cubic splines fit even better. A reasonable model is obtained by having one node at the age of 15.0 years. The basic functions accordingly are defined by (15.27) with $t_0 = 15.0$. All these functions are 0 for $t = 15$ years, so the intercept refers to height at this age. The parameter estimates are given in Table 15.10.

Table 15.10: Cubic spline growth model for 12–17-year-old children with retarded growth.

Fixed effect	Coefficient	S.E.
γ_{00} Intercept	150.00	0.42
$\gamma_{10} f_1$ (linear)	6.43	0.19
$\gamma_{20} f_2$ (quadratic)	0.25	0.13
$\gamma_{30} f_3$ (cubic to the left of 15)	-0.038	0.030
$\gamma_{40} f_4$ (cubic to the right of 15)	-0.529	0.096
Random effect	Variance	S.E.
<i>Level-two (i.e., individual) random effects:</i>		
τ_0^2 Intercept variance	52.07	4.46
τ_1^2 Slope variance f_1	6.23	0.71
τ_2^2 Slope variance f_2	2.59	0.34
τ_3^2 Slope variance f_3	0.136	0.020
τ_4^2 Slope variance f_4	0.824	0.159
<i>Level-one (i.e., occasion) variance:</i>		
σ^2 Residual variance	0.288	0.014
Deviance	6,999.06	

The correlation matrix of the level-two random effects ($U_{0i}, U_{1i}, \dots, U_{4i}$) is estimated as

$$\widehat{R}_U = \begin{pmatrix} 1.0 & 0.26 & -0.31 & 0.32 & 0.01 \\ 0.26 & 1.0 & 0.45 & -0.08 & -0.82 \\ -0.31 & 0.45 & 1.0 & -0.89 & -0.71 \\ 0.32 & -0.08 & -0.89 & 1.0 & 0.40 \\ 0.01 & -0.82 & -0.71 & 0.40 & 1.0 \end{pmatrix}.$$

Testing the random effects (not reported here) shows that the functions f_1, \dots, f_4 all have significant random effects. The level-one residual standard deviation is only $\hat{\sigma} = \sqrt{0.288} = 0.54$ cm, which demonstrates that this family of functions fits rather closely to the height measurements. Notation is made more transparent by defining

$$(t - a)_- = \begin{cases} -(t - a) & (t < a) \\ 0 & (t \geq a), \end{cases}$$

$$(t - a)_+ = \begin{cases} 0 & (t \leq a) \\ t - a & (t > a). \end{cases}$$

Thus, we denote $f_3(t) = (t - 15)_-^3$, $f_4(t) = (t - 15)_+^3$. The mean height curve can be obtained by filling in the estimated fixed coefficients, which yields

$$\begin{aligned} \mathcal{E}(F_i(t)) &= 150.00 + 6.43(t - 15) + 0.25(t - 15)^2 \\ &\quad - 0.038(t - 15)_-^2 - 0.529(t - 15)_+^3. \end{aligned}$$

This function can be differentiated to yield the mean growth rate. Note that $df_3(t)/dt = -3(t-15)_-^2$ and $df_4(t)/dt = 3(t-15)_+^2$. This implies that the mean growth rate is estimated as

$$\text{mean growth rate} = 6.43 + 0.25(t-15) + 0.114(t-15)_-^2 - 1.587(t-15)_+^2.$$

For example, for the minimum of the age range considered, $t = 12$ years, this is $6.43 - (0.25 \times 3) + (0.114 \times 9) + 0 = 6.71$ cm/year. This is slightly larger than the mean growth rate found in the preceding examples for ages from 5 to 10 years. For the maximum age in this range, $t = 17$ years, on the other hand, the average growth rate is $6.43 + (0.25 \times 2) + 0 - (1.587 \times 4) = 0.58$ cm/year, indicating that growth has almost stopped at this age.

The results of this model are illustrated more clearly by a graph. Figure 15.1 presents the average growth curve and a sample of 15 random curves from the population defined by Table 15.10 and the given correlation matrix. The average growth curve does not deviate noticeably from a linear curve for ages below 16 years. It levels off after 16 years. Some of the randomly drawn growth curves are decreasing in the upper part of the range. This is an obvious impossibility for the real growth curves, and indicates that the model is not completely satisfactory for the upper part of this age range. This may be related to the fact that the number of measurements is rather low at ages over 16.5 years.

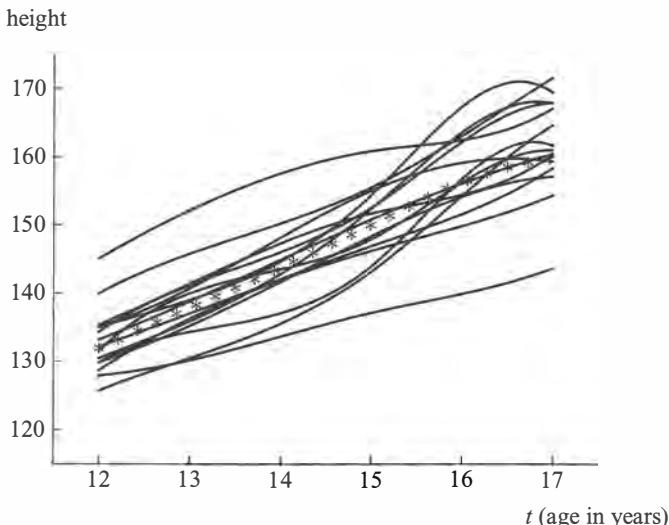


Figure 15.1: Average growth curve (*) and 15 random growth curves for 12–17-year-olds for cubic spline model.

15.2.3 Explaining the functions

Whether one has been using polynomial, spline, or other functions, the random function model can be represented by (15.21), repeated here:

$$F_i(t) = \beta_{0i} + \sum_{h=1}^r \beta_{hi} f_h(t).$$

The same can now be done as in Chapter 5 (see p. 83): the individual-dependent coefficients β_{hi} can be explained with individual-level (i.e., level-two) variables. Suppose there are q

individual-level variables, and denote these variables by Z_1, \dots, Z_q . The inter-individual model to explain the coefficients $\beta_{0i}, \beta_{1i}, \dots, \beta_{ri}$ is then

$$\beta_{hi} = \gamma_{h0} + \gamma_{h1} z_{1i} + \dots + \gamma_{hq} z_{qi} + U_{hi}. \quad (15.29)$$

Substitution of (15.29) into (15.21) yields

$$Y_{ti} = \gamma_{00} + \sum_{h=1}^p \gamma_{h0} f_h(t) + \sum_{k=1}^q \gamma_{0k} z_{ki} + \sum_{k=1}^q \sum_{h=1}^p \gamma_{hk} z_{ki} f_h(t) \\ + U_{0j} + \sum_{h=1}^p U_{hj} f_h(t) + R_{ij}. \quad (15.30)$$

We see that cross-level interactions here are interactions between individual-dependent variables and functions of time. The same approach can be followed as in Chapter 5. The reader may note that this model selection approach, where first a random function model is constructed and then the individual-based coefficients are approached as dependent variables in regression models at level two, is just what was described in Section 6.4.1 as ‘working upward from level one’.

Example 15.11 Explaining growth by gender and height of parents.

We continue the analysis of the retarded growth data of the 12–17-year-olds, and now include the child’s gender and the mean height of the child’s parents. Omitting children with missing value for mother’s or father’s height left 321 children with a total of 1,941 measurements. Gender is coded as +1 for girls and -1 for boys, so that the other parameters give values which are averages over the sexes. Parents’ height is defined as the average height of father and mother minus 165 cm (this value is approximately the mean of the heights of the parents). For parents’ height, the main effect and the interaction effect with f_1 (age minus 15) was included. For gender, the main effect was included as well as the interactions with f_1 and f_2 . These choices were made on the basis of preliminary model fits. The resulting parameter estimates are in Table 15.11.

The correlation matrix of the level-two random effects ($U_{0i}, U_{1i}, \dots, U_{4i}$) is estimated as

$$\widehat{R}_U = \begin{pmatrix} 1.0 & 0.22 & -0.38 & 0.35 & 0.07 \\ 0.22 & 1.0 & 0.38 & -0.05 & -0.81 \\ -0.38 & 0.38 & 1.0 & -0.91 & -0.75 \\ 0.35 & -0.05 & -0.91 & 1.0 & 0.48 \\ 0.07 & -0.81 & -0.75 & 0.48 & 1.0 \end{pmatrix}.$$

The fixed effect of gender (γ_{01}) is not significant, which means that at 15 years there is not a significant difference in height of boys and girls in this population with retarded growth. However, the interaction effects of gender with age, γ_{11} and γ_{21} , show that girls and boys do grow in different patterns during adolescence. The coding of gender implies that the average height difference between girls and boys is given by

$$2(\gamma_{01} + \gamma_{11}f_1(t) + \gamma_{21}f_2(t)) = -0.77 - 2.532(t - 15) - 0.724(t - 15)^2.$$

The girl-boy difference in average growth rate, which is the derivative of this function, is equal to $-2.532 - 1.448(t - 15)$. This shows that from about the age of 13 years (more precisely, for $t > 13.25$), the growth of girls is on average slower than that of boys, and faster before this age.

Parents’ height has a strong main effect: for each centimeter of extra height of the parents, the children are on average 0.263 cm taller at 15 years of age. Moreover, for every centimeter extra of the parents, on average the children grow faster by 0.03 cm/year.

Table 15.11: Growth variability of 12–17-year-old children explained by gender and parents’ height.

Fixed effect	Coefficient	S.E.
γ_{00} Intercept	150.20	0.47
$\gamma_{10} f_1$ (linear)	5.85	0.18
$\gamma_{20} f_2$ (quadratic)	0.053	0.124
$\gamma_{30} f_3$ (cubic to the left of 15)	-0.029	0.030
$\gamma_{40} f_4$ (cubic to the right of 15)	-0.553	0.094
γ_{01} Gender	-0.385	0.426
$\gamma_{11} f_1 \times$ gender	-1.266	0.116
$\gamma_{21} f_2 \times$ gender	-0.362	0.037
γ_{02} Parents’ height	0.263	0.071
$\gamma_{12} f_1 \times$ parents’ height	0.0307	0.0152
Random effect	Variance	S.E.
<i>Level-two (i.e., individual) random effects:</i>		
τ_0^2 Intercept variance	49.71	4.31
τ_1^2 Slope variance f_1	4.52	0.52
τ_2^2 Slope variance f_2	2.37	0.33
τ_3^2 Slope variance f_3	0.132	0.020
τ_4^2 Slope variance f_4	0.860	0.156
<i>Level-one (i.e., occasion) variance:</i>		
σ^2 Residual variance	0.288	0.013
Deviance	6,885.18	

The intercept variance and the slope variances of f_1 and f_2 have decreased, compared to Table 15.10. The residual variance at level one has remained the same, which is natural since the effects included explain differences between curves and do not yield better-fitting curves. The deviance went down by 113.88 points ($df = 5, p < 0.0001$).

15.2.4 Changing covariates

Individual-level variables such as the Z_k of the preceding section are referred to as *constant covariates* because they are constant over time. It is also possible that *changing covariates* are available – changing social or economic circumstances, performance on tests, mood variables, etc. Fixed effects of such variables can be added to the model without problems, but the neat forward model selection approach is disturbed because the changing covariate normally will not be a linear combination of the functions f_h in (15.21). Depending on, for example, the primacy of the changing covariates in the research question, one can employ the changing covariates in the fixed part right from the start (i.e., add this

fixed effect to (15.19) or (15.22)), or incorporate them in the model at the point where the constant covariates are also considered (add the fixed effect to (15.30)).

Example 15.12 *Cortisol levels in infants.*

This example reanalyzes data collected by de Weerth (1998, Chapter 3) in a study on stress in infants; also see de Weerth and van Geert (2002). In this example the focus is not on discovering the shape of the development curves, but on testing a hypothesized effect, while taking into account the longitudinal design of the study. Because of the complexity of the data analysis, we shall combine techniques described in various of the preceding chapters.

The purpose of the study was to investigate experimentally the effect of a stressful event on experienced stress as measured by the cortisol hormone, and whether this effect is stronger in certain hypothesized ‘regression weeks’. The experimental subjects were 18 normal infants with gestational ages ranging from 17 to 37 weeks. Gestational age is age counted from the due day on which the baby should have been born after a complete pregnancy. Each infant was seen repeatedly, the number of visits per infant ranging from 8 to 13, with a total of 222 visits. The infants were divided randomly into an experimental group (14 subjects) and a control group (4 subjects). Cortisol was measured from saliva samples. The researcher collected a saliva sample at the start of the session; then a play session between the mother and the infant followed. For the experimental group, during this session the mother suddenly put the infant down and left the room. In the control group, the mother stayed with the child. After the session, a saliva sample was taken again. This provided a pretest and a post-test measure of the infant’s cortisol level. In this example we consider only the effect of the experiment, not the effect of the ‘regression weeks’. Further information about the study and about the underlying theories can be found in de Weerth (1998, Chapter 3).

The post-test cortisol level was the dependent variable. Explanatory variables were the pretest cortisol level, the group (coded $Z = 1$ for infants in the experimental group and $Z = 0$ for the control group), and gestational age. Gestational age is the time variable t . In order to avoid very small coefficients, the unit is chosen as 10 weeks, so that t varies between 1.7 and 3.7. A preliminary inspection of the joint distribution of pretest and post-test showed that cortisol levels had a rather skewed distribution (as is usual for cortisol measurements). This skewness was reduced satisfactorily by using the square root of the cortisol level, both for the pretest and for the post-test. These variables are denoted by X and Y , respectively. The pretest variable, X , is a changing covariate.

The law of initial value proposed by Wilder in 1956 implies that an inverse relation is expected between basal cortisol level and the subsequent response to a stressor (cf. de Weerth, 1998, p. 41). An infant with a low basal level of cortisol will accordingly react to a stressor with a cortisol increase, whereas an infant with a high basal cortisol level will react to a stressor with a cortisol decrease. The basal cortisol level is measured here by the pretest. The average of X (the square root of the pretest cortisol value) was 2.80. This implies that infants with a basal value x less than about 2.80 are expected to react to stress with a relatively high value for Y whereas infants with x more than about 2.80 are expected to react to stress with a relatively low value of Y . The play session itself may also lead to a change in cortisol value, so there may be a systematic difference between X and Y . Therefore, the stress reaction was investigated by testing whether the experimental group has a more strongly negative regression coefficient of Y on $X - 2.80$ than the control group; in other words, the research hypothesis is that there is a negative interaction between Z and $X - 2.80$ in their effect on Y .

It appeared that a reasonable first model for the square root of the post-test cortisol level, if the difference between the experimental and control group is not yet taken into account, is the model where the changing covariate, defined as $X - 2.80$, has a fixed effect, while time (i.e., gestational age) has a linear fixed as well as random effect:

$$Y_{ti} = \gamma_{00} + \gamma_{10} t + \gamma_{01} (x_{ti} - 2.80) + U_{0i} + U_{1i} t + R_{ti}.$$

The results are given as Model 1 in Table 15.12. This model can be regarded as the null hypothesis model, against which we shall test the hypothesis of the stressor effect.

Table 15.12: Estimates for two models for cortisol data.

Fixed effect	Model 1		Model 2	
	Coefficient	S.E.	Coefficient	S.E.
γ_{00} Intercept	3.21	0.33	3.16	0.34
γ_{10} Gestational age	-0.201	0.117	-0.185	0.118
$\gamma_{01} X - 2.80$	0.358	0.045	0.532	0.096
$\gamma_{02} Z$			0.009	0.117
$\gamma_{03} Z \times (X - 2.80)$			-0.224	0.108
Random effect	Parameter	S.E.	Parameter	S.E.
<i>Level-two (i.e., individual) random effects:</i>				
τ_0^2 Intercept variance	1.25	0.66	1.32	0.68
τ_1^2 Slope variance	0.151	0.082	0.155	0.082
τ_{01} Intercept-slope covariance	-0.43	0.23	-0.45	0.23
<i>Level-one (i.e., occasion) variance:</i>				
σ^2 Residual variance	0.175	0.018	0.172	0.018
Deviance	280.57		276.42	

The random effect of gestational age is significant. The model without this effect, not shown in the table, has deviance 289.46, so that the deviance comparison yields $\chi^2 = 8.89$, $df = 2$. Using the mixture of chi-squared distributions with 1 and 2 df , according to Section 6.2.1 and Table 6.2, we obtain $p < 0.05$. The pretest has a strongly significant effect ($t = 8.0$, $p < 0.0001$): children with a higher basal cortisol value tend also to have a higher post-test cortisol value. The fixed effect of gestational age is not significant ($t = 1.67$, two-sided $p > 0.05$). The intercept variance is rather large because the time variable is not centered and $t = 0$ refers to the due date of birth, quite an extrapolation from the sample data.

To test the effect of the stressor, the fixed effect of the experimental group Z and the interaction effect of $Z \times (X - 2.80)$ were added to the model. The theory, that is, Wilder's law of initial value, predicts the effect of the product variable $Z \times (X - 2.80)$ and therefore the test is focused on this effect. The main effect of Z is included only because including an interaction effect without the corresponding main effects can lead to errors of interpretation.

The result is presented as Model 2 in Table 15.12. The stressor effect (parameter γ_{03} in the table) is significant ($t = -2.07$ with many degrees of freedom because this is the effect of a level-one variable, one-sided $p < 0.025$). This confirms the hypothesized effect of the stressor in the experimental group.

The model fit is not good, however. The standardized level-two residuals defined by (10.10) were calculated for the 18 infants. For the second infant ($i = 2$) the value was $S_i^2 = 34.76$ ($df = n_i = 13$, $p = 0.0009$). With the Bonferroni correction which takes into account that this is the most significant out of 18 values, the significance value still is $18 \times 0.0009 = 0.016$. Inspection of the data showed that this was a child who had been asleep shortly before many of the play sessions. Being just awake is known to have a potential effect on cortisol values. Therefore, for each session it had

been recorded whether the child had been asleep in the half hour immediately preceding the session. Subsequent data analysis showed that having been asleep (represented by a dummy variable equal to 1 if the infant had been asleep in the half hour preceding the session and equal to 0 otherwise) had an important fixed effect, not a random effect, at level two, and also was associated with heteroscedasticity at level one (see Chapter 8). Including these two effects led to the estimates presented in Table 15.13.

Table 15.13: Estimates for a model controlling for having been asleep.

Model 3		
Fixed effect	Coefficient	S.E.
γ_{00} Intercept	3.04	0.31
γ_{10} Gestational age	-0.142	0.099
$\gamma_{01} X - 2.80$	0.548	0.084
$\gamma_{02} Z$	-0.058	0.118
$\gamma_{03} Z \times (X - 2.80)$	-0.136	0.096
γ_{04} Sleeping	0.358	0.117
Random effect	Parameter	S.E.
<i>Level-two (i.e., individual) random effects:</i>		
τ_0^2 Intercept variance	0.869	0.496
τ_1^2 Slope variance	0.092	0.057
τ_{01} Intercept-slope covariance	-0.28	0.17
<i>Level-one (i.e., occasion) variance parameters:</i>		
σ_0^2 Basic residual variance	0.130	0.015
σ_{01} Sleeping effect	0.116	0.045
Deviance	251.31	

The two effects of sleeping are jointly strongly significant (the comparison between the deviances of Models 2 and 3 yields $\chi^2 = 25.11$, $df = 2$, $p < 0.0001$). The estimated level-one variance is 0.130 for children who had not slept and (using formula (8.1)) 0.362 for children who had slept. But the stressor ($Z \times (X - 2.80)$ interaction) effect has now lost its significance ($t = -1.42$, n.s.). The most significant standardized level-two residual defined by (10.10) for this model is obtained for the fourth child ($j = 4$), with the value $S_j^2 = 29.86$, $df = n_j = 13$, $p = 0.0049$. Although this is a rather small p -value, the Bonferroni correction now leads to a significance probability of $18 \times 0.0049 = 0.09$, which is not alarmingly low. The fit of this model therefore seems satisfactory, and the estimates do not support the hypothesized stressor effect. However, these results cannot be interpreted as evidence *against* the stressor effect, because the parameter estimate does have the predicted negative sign and the number of experimental subjects is not very large, so that the power may have been low.

It can be concluded that it is important to control for the infant having slept shortly before the play session, and in this case this control makes the difference between a significant and a nonsignificant effect. Having slept not only leads to a higher post-test cortisol value, controlling for the pretest value, but also triples the residual variance at the occasion level.

15.3 Autocorrelated residuals

Of the relevant extensions to the hierarchical linear model, we wish briefly to mention the hierarchical linear model with *autocorrelated residuals*. In the fixed occasion design, the assumption that the level-one residuals R_{ti} are independent can be relaxed and replaced by the assumption of first-order autocorrelation,

$$R_{1i} = R_{1,i}^{(0)}, \quad R_{t+1,i} = \rho R_{ti} + \sqrt{1 - \rho^2} R_{t+1,i}^{(0)} \quad (t \geq 1) \quad (15.31)$$

where $R_{t,i}^{(0)}$ are independent and identically distributed variables. The parameter ρ is called the *autocorrelation coefficient*. This model represents a type of dependence between adjacent observations, which quickly dies out between observations which are further apart. The correlation between the residuals is

$$\rho(R_{ti}, R_{sj}) = \rho^{|t-s|}. \quad (15.32)$$

Other covariance and correlation patterns are also possible. For variable occasion designs, level-one residuals with the correlation structure (15.32) also are called autocorrelated residuals, although they cannot be constructed by the relations (15.31). These models are discussed in various textbooks, such as Diggle et al. (2002), Goldstein (2011, Section 5.4), Verbeke and Molenbergs (2000, Chapter 10), Singer and Willett (2003, Section 7.3), and Hedeker and Gibbons (2006).

The extent to which time dependence can be modeled by random slopes, or rather by autocorrelated residuals, or a combination of both, depends on the phenomenon being modeled. This issue usually will have to be decided empirically; most computer programs implementing the hierarchical linear model allow the inclusion of autocorrelated residuals in the model.

15.4 Glommary

Longitudinal data. Repeated measurements of the same characteristic for a sample of individual subjects. Another name is *panel data*. The level-one units here are the measurements (usually labeled as time points) and the level-two units are individual subjects or respondents.

Fixed occasion designs. Data structures with a fixed set of measurement occasions for all subjects, which may be conveniently labeled $t = 1, \dots, m$. This does not exclude that data may be incomplete for some or all of the subjects.

***t*-test for paired samples.** This provides the simplest case of a longitudinal design with fixed measurement occasions.

Random intercept model for repeated measures. This is also called the *compound symmetry model*.

Random slope models. These models, with random slopes for time and transformed time variables, can be used to represent time-dependent variances as well as nonconstant correlations between different measurements of the same subject.

Complete data vector. This can be defined for fixed measurement occasions as the vector of data for all occasions. In this case the specification of the random part can be regarded as the way to define assumptions about the covariance matrix of the complete data vector.

Fully multivariate model. This is the model in which the covariance matrix of the complete data vector is completely unrestricted. This model can be used for longitudinal data with a fixed occasion design, but also for multivariate regression analysis where the components of the dependent variable are not necessarily repeated measures of the same characteristic.

Proportion of explained variance in longitudinal models. This can be defined as the proportional reduction in prediction error for individual measurements, averaged over the m measurement occasions.

Variable occasion designs. In these designs the data are ordered according to some underlying dimension such as time, and for each individual data are recorded at some set of time points which is not necessarily related to the time points at which the other individuals are observed. Repeated measures according to variable occasion designs can be regarded as observations on a population of curves.

Multilevel approach to populations of curves. This approach can represent the average curve by fixed effects of individual-level variables and of their cross-level interactions with time variables. Individual deviations are represented by random effects of time variables. In addition, changing covariates can be included to represent further individual deviations.

Families of curves. Various families of curves were proposed for modeling such a population of curves: polynomial functions, spline functions, and piecewise linear functions.

Spline functions. Smooth piecewise polynomials. Splines are polynomial functions between points called *nodes*, the polynomials being smoothly linked at the nodes. Denoting the degree of the piecewise polynomials by p , for $p = 1$ we obtain piecewise linear functions, for $p = 2$ quadratic splines, and for $p = 3$ cubic splines. Spline functions are generally more flexible in use than polynomial functions.

Autocorrelated residuals. A further way to represent correlation patterns of residuals across time.

16

Multivariate Multilevel Models

This chapter is devoted to the multivariate version of the hierarchical linear model treated in Chapters 4 and 5. The term ‘multivariate’ refers here to the dependent variable: there are assumed to be two or more dependent variables. The model of this chapter is a three-level model, with variables as level-one units, individuals as level-two units, and groups as level-three units. One example is the nesting structure of Chapter 15, measurements nested within individuals, combined with the nesting structure considered in the earlier chapters, individuals within groups. This leads to three levels: repeated measurements within individuals within groups. The model of this chapter, then, is an extension of the fully multivariate model of Section 15.1.3 to the case where individuals are nested in groups. However, in the present chapter, the dependent variables are not necessarily supposed to be longitudinal measurements of the same characteristic (although they could be).

As an example of multivariate multilevel data, think of pupils (level-two units) i in classes (level-three units) j , with m variables, Y_1, \dots, Y_m , being measured for each pupil if data are complete. The measurements are the level-one units and could refer, for example, to test scores in various school subjects but also variables measured in completely unrelated and incommensurable scales such as attitude measurements.

The dependent variable is denoted

Y_{hij} is the measurement on the h ’th variable
for individual i in group j .

It is not necessary that, for each individual i in each group j , an observation of each of the m variables is available. (It must be assumed, however, that missingness is at random, that is, the availability of a measurement should be unrelated to its residual; cf. Chapter 9).

Just as in Chapter 15, the complete data vector can be defined as the vector of data for the individual, possibly hypothetical, who does have observations on all variables. This complete data vector is denoted by

$$Y^c = \begin{pmatrix} Y_{1ij} \\ \vdots \\ Y_{mij} \end{pmatrix}.$$

If a researcher is interested in more than one dependent variable, it may still not be necessary to analyze them simultaneously as a multivariate dependent variable. Therefore, by way of introduction, we discuss some arguments for making this additional effort. Subsequently, we discuss the multivariate random intercept model and, briefly, the multivariate random slope model.

16.1 Why analyze multiple dependent variables simultaneously?

It is possible to analyze all m dependent variables separately. There are several reasons why it may be sensible to analyze the data jointly, that is, as multivariate data.

1. Conclusions can be drawn about the correlations between the dependent variables – notably, the extent to which the unexplained correlations depend on the individual and on the group level. Such conclusions follow from the partitioning of the covariances between the dependent variables over the levels of analysis.
2. The tests of specific effects for single dependent variables are more powerful in the multivariate analysis. This will be visible in the form of smaller standard errors. The additional power is negligible if the dependent variables are only weakly correlated, but may be considerable if the dependent variables are strongly correlated while at the same time the data are very incomplete, that is, the average number of measurements available per individual is considerably less than m .
3. Testing whether the effect of an explanatory variable on dependent variable Y_1 is larger than its effect on Y_2 , when the data on Y_1 and Y_2 were observed (totally or partially) on the same individuals, is possible only by means of a multivariate analysis.
4. If one wishes to carry out a single test of the joint effect of an explanatory variable on several dependent variables, then a multivariate analysis is also required. Such a single test can be useful, for example, to avoid the danger of capitalization on chance which is inherent in carrying out a separate test for each dependent variable.

A multivariate analysis is more complicated than separate analyses for each dependent variable. Therefore, when one wishes to analyze several dependent variables, the greater complexity of the multivariate analysis will have to be balanced against the reasons listed above. Often it is advisable to start by analyzing the data for each dependent variable separately.

16.2 The multivariate random intercept model

Suppose there are covariates X_1, \dots, X_p , which may be individual-dependent or group-dependent. The random intercept model for dependent variable Y_h is expressed by the formula

$$Y_{hij} = \gamma_{0h} + \gamma_{1h} x_{1ij} + \gamma_{2h} x_{2ij} + \dots + \gamma_{ph} x_{pij} + U_{hj} + R_{hij}. \quad (16.1)$$

In words: for the h th dependent variable, the intercept is γ_{0h} , the regression coefficient on X_1 is γ_{1h} , the coefficient on X_2 is γ_{2h} , ..., the random part of the intercept in group j is U_{hj} , and the residual is R_{hij} . This is just a random intercept model like (4.5) and (4.9). Since the variables Y_1, \dots, Y_m are measured on the same individuals, however, their dependence can be taken into account. In other words, the U s and R s are regarded as components of vectors

$$R_{ij} = \begin{pmatrix} R_{1ij} \\ \vdots \\ R_{mij} \end{pmatrix}, \quad U_j = \begin{pmatrix} U_{1j} \\ \vdots \\ U_{mj} \end{pmatrix}.$$

Instead of residual variances at levels one and two, there are now residual covariance matrices,

$$\Sigma = \text{cov}(R_{ij}) \quad \text{and} \quad T = \text{cov}(U_j).$$

The covariance matrix of the complete observations, conditional on the explanatory variables, is the sum of these,

$$\text{var}(Y^c) = \Sigma + T. \quad (16.2)$$

To represent the multivariate data in the multilevel approach, three nesting levels are used. The first level is that of the dependent variables indexed by $h = 1, \dots, m$, the second level is that of the individuals $i = 1, \dots, n_j$, and the third level is that of the groups, $j = 1, \dots, N$. So each measurement of a dependent variable on some individual is represented by a separate line in the data matrix, containing the values $i, j, h, Y_{hij}, x_{1ij}$, and those of the other explanatory variables.

The multivariate model is formulated as hierarchical linear model using the same trick as in Section 15.1.3. Dummy variables d_1, \dots, d_m are used to indicate the dependent variables, just as in formula (15.2). Dummy variable d_h is 1 or 0, depending on whether the data line refers to dependent variable Y_h or to one of the other dependent variables. Formally, this is expressed by

$$d_{shij} = \begin{cases} 1 & (h = s), \\ 0 & (h \neq s). \end{cases} \quad (16.3)$$

With these dummies, the random intercept models (16.1) for the m dependent variables can be integrated into one three-level hierarchical linear model by the expression

$$Y_{hij} = \sum_{s=1}^m \gamma_{0s} d_{shij} + \sum_{k=1}^p \sum_{s=1}^m \gamma_{ks} d_{shij} x_{kij} + \sum_{s=1}^m U_{sj} d_{shij} + \sum_{s=1}^m R_{sij} d_{shij}. \quad (16.4)$$

All variables (including the constant) are multiplied by the dummy variables. Note that the definition of the dummy variables implies that in the sums over $s = 1, \dots, m$ only the term for $s = h$ gives a contribution and all other terms disappear. So this formula is just a complicated way of rewriting formula (16.1).

The purpose of this formula is that it can be used to obtain a multivariate hierarchical linear model. The variable-dependent random residuals R_{hij} in this formula are random

slopes at level two of the dummy variables, R_{sij} being the random slope of d_s , and the random intercepts U_{hj} become the random slopes at level three of the dummy variables. There is no random part at level one.

This model can be further specified, for example, by omitting some of the variables X_k from the explanation of some of the dependent variables Y_h . This amounts to dropping some of the terms $\gamma_{ks} d_{shij} x_{kij}$ from (16.4). Another possibility is to include variable-specific covariates, in analogy to the changing covariates of Section 15.2.4. An example of this is a study of school pupils' performance on several subjects (these are the multiple dependent variables), using the pupils' motivation for each of the subjects separately as explanatory variables.

Multivariate empty model

The multivariate empty model is the multivariate model without explanatory variables. For $m = 2$ variables, this is just the model of Section 3.6.1. Writing out formulas (16.1) and (16.4) without explanatory variables (i.e., with $p = 0$) yields the following formula, which is the specification of the three-level multivariate empty model:

$$\begin{aligned} Y_{hij} &= \gamma_{0h} + U_{hj} + R_{hij} \\ &= \sum_{s=1}^m \gamma_{0s} d_{shij} + \sum_{s=1}^m U_{sj} d_{shij} + \sum_{s=1}^m R_{sij} d_{shij}. \end{aligned} \quad (16.5)$$

This empty model can be used to decompose the raw variances and covariances into parts at the two levels. When referring to the multivariate empty model, the covariance matrix

$$\Sigma = \text{cov}(R_{ij})$$

may be called the *within-group covariance matrix*, while

$$T = \text{cov}(U_j)$$

may be called the *between-group covariance matrix*. In the terminology of Chapter 3, these are the *population* (and not *observed*) covariance matrices. In Chapter 3 we saw that, if the group sizes n_j all are equal to n , then the covariance matrix between the group means is

$$T + \frac{1}{n} \Sigma \quad (16.6)$$

(cf. equation (3.9)).

Section 3.6.1 presented a relatively simple way to estimate the within- and between-group correlations from the intraclass coefficients and the observed total and within-group correlations. Another way is to fit the multivariate empty model (16.5) using multilevel software by the ML or the REML method. For large sample sizes these methods will provide virtually the same results, but for small sample sizes the ML and REML methods will provide more precise estimators. The gain in precision will be especially large if the correlations are high while there are relatively many missing data (i.e., there are many individuals who provide less than m dependent variables, but missingness is still at random).

Example 16.1 Language and arithmetic scores in elementary schools.

The example used throughout Chapters 4 and 5 is extended by analyzing not only the scores on the language test but also those on the arithmetic test. So there are $m = 2$ dependent variables. In the multivariate model there may be missing values on one of the dependent variables, but not both, because each data point must make some contribution to the dependent variables. We now use a version of the data set containing 3,773 pupils, 4 of whom are missing an arithmetic test score and 7 of whom are missing a language score.

Table 16.1: Parameter estimates for multivariate empty model.

	Language $h = 1$		Arithmetic $h = 2$		(Covariance)	
	Par.	S.E.	Par.	S.E.	Par.	S.E.
Fixed effect						
γ_{0h} Intercept	40.98	0.32	19.44	0.26		
Random effect	Par.	S.E.	Par.	S.E.	Par.	S.E.
$\tau_h^2 = \text{var}(U_{hj})$	17.86	2.13	12.45	1.41		
$\tau_{12} = \text{cov}(U_{1j}, U_{2j})$					13.77	1.63
<i>Within-schools covariance matrix:</i>						
$\sigma_h^2 = \text{var}(R_{hij})$	62.87	1.49	32.12	0.76		
$\sigma_{12} = \text{cov}(R_{1ij}, R_{2ij})$					28.51	0.89
Deviance			48,708.7			

First the multivariate empty model, represented by (16.5), is fitted. The results are given in Table 16.1. The results for the language test may be compared with the univariate empty model for the language test, presented in Table 4.1 of Chapter 4. The parameter estimates are slightly different, which is a consequence of the fact that the estimation procedure is now multivariate. The extra result of the multivariate approach is the estimated covariance at level three, $\text{cov}(U_{1j}, U_{2j}) = 13.77$, and at level two, $\text{cov}(R_{1ij}, R_{2ij}) = 28.51$. The corresponding population correlation coefficients at the school and pupil level are, respectively,

$$\rho(U_{1j}, U_{2j}) = \frac{13.77}{\sqrt{12.45 \times 17.86}} = 0.92,$$

$$\rho(R_{1ij}, R_{2ij}) = \frac{28.51}{\sqrt{32.12 \times 62.87}} = 0.63.$$

This shows that especially the random school effects for language and arithmetic are very strongly correlated.

For the correlations between observed variables, these estimates yield a correlation between individuals of (cf. (16.2))

$$\hat{\rho}(Y_{1ij}, Y_{2ij}) = \frac{13.77 + 28.51}{\sqrt{(17.86 + 62.87)(12.45 + 32.12)}} = 0.66$$

Table 16.2: Parameter estimates for multivariate model for language and arithmetic tests.

Fixed effect	Language $h = 1$		Arithmetic $h = 2$		(Covariance)	
	Par.	S.E.	Par.	S.E.	Par.	S.E.
γ_{0h} Intercept	41.27	0.22	19.67	0.19		
γ_{1h} IQ	2.210	0.056	1.315	0.044		
γ_{2h} SES	0.174	0.012	0.119	0.009		
γ_{3h} $\overline{\text{IQ}}$	0.894	0.315	1.135	0.267		
γ_{4h} $\overline{\text{SES}}$	-0.098	0.044	-0.062	0.037		
γ_{5h} IQ \times SES	-0.017	0.005	-0.005	0.004		
γ_{6h} $\overline{\text{IQ}} \times \overline{\text{SES}}$	-0.082	0.033	-0.050	0.028		
Random effect	Par.	S.E.	Par.	S.E.	Par.	S.E.
<i>Residual between-schools covariance matrix:</i>						
$\tau_h^2 = \text{var}(U_{hj})$	8.12	1.02	6.17	0.74		
$\tau_{12} = \text{cov}(U_{1j}, U_{2j})$					5.88	0.77
<i>Residual within-schools covariance matrix:</i>						
$\sigma_h^2 = \text{var}(R_{hij})$	37.97	0.90	22.92	0.54		
$\sigma_{12} = \text{cov}(R_{1ij}, R_{2ij})$					13.42	0.54
Deviance			46,564.4			

and, for groups of a hypothetical size $n = 30$, a correlation between group means (cf. (16.6)) of

$$\hat{\rho}(\bar{Y}_{1j}, \bar{Y}_{2j}) = \frac{13.77 + 28.51/30}{\sqrt{(17.86 + 62.87/30)(12.45 + 32.12/30)}} = 0.90.$$

Explanatory variables included are IQ, SES, the group mean of IQ, and the group mean of SES. As in the examples in Chapters 4 and 5, the IQ measurement is the verbal IQ from the ISI test. The correspondence with formulas (16.1) and (16.4) is that X_1 is IQ, X_2 is SES, X_3 is the group mean of IQ, X_4 is the group mean of SES, $X_5 = X_1 \times X_2$ represents the interaction between IQ and SES, and $X_6 = X_3 \times X_4$ represents the interaction between group mean IQ and group mean SES. The results are given in Table 16.2.

Calculating t -statistics for the fixed effects shows that for language, all effects are significant at the 0.05 level. For arithmetic the fixed effect of mean SES is not significant, nor are the two interaction effects, but the other three fixed effects are significant. The residual correlations are $\rho(U_{1j}, U_{2j}) = 0.83$ at the school level and $\rho(R_{1ij}, R_{2ij}) = 0.45$ at the pupil level. This shows that taking the explanatory variables into account has led to somewhat smaller, but still substantial residual correlations. Especially the school-level residual correlation is large. This suggests that, also when controlling for IQ, SES, mean IQ, and mean SES, the factors at school level that determine language and arithmetic proficiency are the same. Such factors could be associated with school policy but also with aggregated pupil characteristics not taken into account here, such as average performal IQ.

When the interaction effect of mean IQ with mean SES is to be tested for both dependent variables simultaneously, this can be done by fitting the model from which these interaction effects are excluded. In formula (16.1) this corresponds to the effects γ_{61} and γ_{62} of X_6 on Y_1 and Y_2 ; in formula

(16.4) this corresponds to the effects γ_{61} and γ_{62} of d_1X_6 and d_2X_6 . The model from which these effects are excluded has a deviance of 46,570.5, which is 6.2 less than the model of Table 16.2. In a chi-squared distribution with $df = 2$, this is a significant result ($p < 0.05$).

16.3 Multivariate random slope models

The notation is quite complex already, and therefore only the case of one random slope is treated here. More random slopes are, in principle, a straightforward extension.

Suppose that variable X_1 has a random slope for the various dependent variables. For the h th dependent variable, denote the random intercept by U_{0hj} and the random slope of X_1 by U_{1hj} . The model for the h th dependent variable is then

$$Y_{hij} = \gamma_{0h} + \gamma_{1h}x_{1ij} + \dots + \gamma_{ph}x_{pij} + U_{0hj} + U_{1hj}x_{1ij} + R_{hij}. \quad (16.7)$$

The random slopes U_{1hj} are uncorrelated between groups j but correlated between variables h ; this correlation between random slopes of different dependent variables is a parameter of this model that is not included in the hierarchical linear model of Chapter 5.

Just like the multilevel random intercept model, this model is implemented by a three-level formulation, defined by

$$Y_{hij} = \sum_{s=1}^m \gamma_{0s} d_{shij} + \sum_{k=1}^p \sum_{s=1}^m \gamma_{ks} d_{shij} x_{kij} + \sum_{s=1}^m U_{0sj} d_{shij} + \sum_{s=1}^m U_{1sj} d_{shij} x_{1ij} + \sum_{s=1}^m R_{sij} d_{shij}. \quad (16.8)$$

This means that again, technically, there is no random part at level one, there are m random slopes at level two (of variables d_1, \dots, d_m) and $2m$ random slopes at level three (of variables d_1, \dots, d_m and of the product variables d_1X_1, \dots, d_mX_1). With this kind of model, an obvious further step is to try and model the random intercepts and slopes by group-dependent variables as in Section 5.2.

16.4 Glommary

Multivariate multilevel model. This is a three-level model, in which level one consists of variables or measurements; level two consists of individuals; and level three of groups.

Dummy variables. To specify this three-level model, dummy variables are used to represent the variables. The dummy variables define the variables represented by the level-one units, and have random slopes at level two (individuals) to represent the between-individual within-group variation, and random slopes at level three (groups) to represent the between-group variation.

Multivariate random intercept model. This represents within-group and between-group variances, covariances, and correlations, as also discussed in Section 3.6.

Multivariate random slope model. This has, in addition, random slopes for dependent variables Y_h as a function of one or more explanatory variables X_1 , etc. These random slopes are correlated between the dependent variables.

17

Discrete Dependent Variables

Up to now, it has been assumed in this book that the dependent variable has a continuous distribution and that the residuals at all levels (U_{0j} , R_{ij} , etc.) have normal distributions. This provides a satisfactory approximation for many data sets. However, there also are many situations where the dependent variable is discrete and cannot be well approximated by a continuous distribution. This chapter treats the hierarchical generalized linear model, which is a multilevel model for discrete dependent variables.¹

OVERVIEW OF THE CHAPTER

After a brief overview of the reasons for employing hierarchical generalized linear models, this chapter first focuses on multilevel modeling for dichotomous outcome variables (with only two values). The basic idea of logistic regression is presented, and then the random intercept and random slope logistic models are treated. It is shown how intraclass correlation coefficients and measures of proportion of explained variance can be defined here in slightly different ways than for continuous outcomes. Subsequently a hierarchical generalized linear model is treated for ordered categorical variables (with a small number of ordered categories, for example, 1, 2, 3, 4); and for counts (with natural numbers as values: 0, 1, 2, 3, ...).

17.1 Hierarchical generalized linear models

Important instances of discrete dependent variables are dichotomous variables (e.g., success versus failure of whatever kind) and counts (e.g., in the study of some kind of event, the number of events happening in a predetermined time period). It is usually unwise to apply linear regression methods to such variables, for two reasons.

The first reason is that the range of such a dependent variable is restricted, and the usual linear regression model might take its fitted value outside this allowed range. For example, dichotomous variables can be represented as having the values 0 and 1. A fitted value of, say, 0.7, can still be interpreted as a probability of 0.7 for outcome 1 and a probability of 0.3 for outcome 0. But what about a fitted value of -0.21 or 1.08 ? A meaningful model for

outcomes that have only the values 0 or 1 should not allow fitted values that are negative or greater than 1. Similarly, a meaningful model for count data should not lead to negative fitted values.

The second reason is of a more technical nature, and is the fact that for discrete variables there is often some natural relation between the mean and the variance of the distribution. For example, for a dichotomous variable Y that has probability p for outcome 1 and probability $1 - p$ for outcome 0, the mean is

$$\mathcal{E}Y = p$$

and the variance is

$$\text{var}(Y) = p(1 - p). \quad (17.1)$$

Thus, the variance is not a free parameter but is determined by the mean; and the variance is not constant: there is heteroscedasticity. In terms of multilevel modeling, this could lead to a relation between the parameters in the fixed part and the parameters in the random part.

This has led to the development of regression-like models that are more complicated than the usual multiple linear regression model and that take account of the nonnormal distribution of the dependent variable, its restricted range, and the relation between mean and variance. The best-known method of this kind is *logistic regression*, a regression-like model for dichotomous data. *Poisson regression* is a similar model for count data. In the statistical literature, such models are known as *generalized linear models*; see McCullagh and Nelder (1989) or Long (1997).

The present chapter gives an introduction to multilevel versions of some generalized linear models; these multilevel versions are aptly called hierarchical generalized linear models or generalized linear mixed models.

17.2 Introduction to multilevel logistic regression

Logistic regression (e.g., Hosmer and Lemeshow, 2000; Long, 1997; McCullagh and Nelder, 1989; Ryan, 1997) is a kind of regression analysis for dichotomous, or binary, outcome variables, that is, outcome variables with two possibly values such as pass/fail, yes/no, or like/dislike. The reader is advised to study an introductory text on logistic regression if he or she is not already acquainted with this technique.

17.2.1 Heterogeneous proportions

The basic data structure of two-level logistic regression is a collection of N groups ('units at level two'), with, in group j ($j = 1, \dots, N$), a random sample of n_j level-one units ('individuals'). The outcome variable is dichotomous and denoted by Y_{ij} for level-one unit i in group j . The two outcomes are coded 0 for 'failure' and 1 for 'success'. The total sample size is denoted by $M = \sum_j n_j$. If one does not (yet) take explanatory variables into account, the probability of success is regarded as constant in each group. The success probability in group j is denoted by P_j . In a random coefficient model (cf. Section 4.3.1), the groups are considered as being taken from a population of groups and the success probabilities in the

groups, P_j , are regarded as random variables defined in this population. The dichotomous outcome can be represented as the sum of this probability and a residual,

$$Y_{ij} = P_j + R_{ij}. \quad (17.2)$$

In words, the outcome for individual i in group j , which is either 0 or 1, is expressed as the sum of the probability (average proportion of successes) in this group plus some individual-dependent residual. This residual has (like all residuals) mean zero but for these dichotomous variables it has the peculiar property that it can assume only the values $-P_j$ and $1 - P_j$, since (17.2) must be 0 or 1. A further peculiar property is the fact that, given the value of the probability P_j , the variance of the residual is

$$\text{var}(R_{ij}) = P_j(1 - P_j), \quad (17.3)$$

in accordance with formula (17.1).

Equation (17.2) is the dichotomous analog of the empty (or unconditional) model defined for continuous outcomes in equations (3.1) and (4.6). Section 3.3.1 remains valid for dichotomous outcome variables, with P_j taking the place of $\mu + U_j$, except for one subtle distinction. In the empty model for continuous outcome variables it was assumed that the level-one residual variance was constant. This is not adequate here because, in view of formula (17.3), the groups have different within-group variances. Therefore the parameter σ^2 must be interpreted here as the *average residual variance*, that is, the average of (17.3) in the population of all groups. With this modification in the interpretation, the formulas of Section 3.3.1 still are valid. For example, the intraclass correlation coefficient is still defined by (3.2) and can be estimated by (3.12). Another definition of the intraclass correlation is also possible, however, as is mentioned below in Section 17.3.2.

Since the outcome variable is coded 0 and 1, the group average

$$\bar{Y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij} \quad (17.4)$$

is now the proportion of successes in group j . This is an estimate of the group-dependent probability P_j . Similarly, the overall average

$$\hat{P}_\cdot = \bar{Y}_\cdot = \frac{1}{M} \sum_{j=1}^N \sum_{i=1}^{n_j} Y_{ij} \quad (17.5)$$

here is the overall proportion of successes.

Testing heterogeneity of proportions

To test whether there are indeed systematic differences between the groups, the well-known chi-squared test can be used. The test statistic of the chi-squared test for a contingency table is often given in the familiar form $\sum(O - E)^2/E$, where O is the observed and E the expected count in a cell of the contingency table. In this case it can be written also as

$$X^2 = \sum_{j=1}^N n_j \frac{(\bar{Y}_j - \hat{P}_\cdot)^2}{\hat{P}_\cdot(1 - \hat{P}_\cdot)}. \quad (17.6)$$

It can be tested against the chi-squared distribution with $N - 1$ degrees of freedom. This chi-squared distribution is an approximation valid if the expected numbers of successes and of failures in each group, $n_j \bar{Y}_j$ and $n_j(1 - \bar{Y}_j)$, respectively, are all at least 1 while 80% of them are at least 5 (cf. Agresti, 2002). This condition will not always be satisfied, and the chi-squared test may then be seriously in error. For a large number of groups the null distribution of X^2 can then be approximated by a normal distribution with the correct mean and variance (Haldane, 1940; McCullagh and Nelder, 1989, p. 244), or an exact permutation test may be used.

Another *test of heterogeneity of proportions* was proposed by Commenges and Jacqmin (1994). The test statistic is

$$T = \frac{\sum_{j=1}^N \left\{ n_j^2 (\bar{Y}_j - \hat{P})^2 \right\} - M \hat{P} (1 - \hat{P})}{\hat{P} (1 - \hat{P}) \sqrt{2 \sum_{j=1}^N n_j (n_j - 1)}}. \quad (17.7)$$

Large values of this statistic are an indication of heterogeneous proportions. This statistic can be tested against a standard normal distribution.

The fact that the numerator contains a weight of n_j^2 whereas the chi-squared test uses the weight n_j shows that these two tests combine the groups in different ways: larger groups count more strongly in the Commenges–Jacqmin test. When the group sizes n_j are different, it is possible that the two tests lead to different outcomes. A similar test for the intercept variance, also controlling for fixed effects, was proposed by Commenges et al. (1994).

The advantage of test (17.7) over the chi-squared test is that, when group sizes differ, it has a higher power to test randomness of the observations against the empty model treated below, that is, against the alternative hypothesis represented by (17.10) with $\tau_0^2 > 0$. A further practical advantage is that it can be applied whenever there are many groups, even with small group sizes, provided that no single group dominates. A rule of thumb for the application of this test is that there should be at least $N = 10$ groups, the biggest group should not have a relative share larger than $n_j/M = 0.10$, and the ratio of the largest group size to the 10th largest group size should not be more than 10.

Estimation of between- and within-groups variance

The true variance between the group-dependent probabilities, that is, the population value of $\text{var}(P_j)$, can be estimated by formula (3.11), repeated here,

$$\hat{\tau}^2 = S_{\text{between}}^2 - \frac{S_{\text{within}}^2}{\tilde{n}},$$

where \tilde{n} is defined as in (3.7) by

$$\tilde{n} = \frac{1}{N-1} \left\{ M - \frac{\sum_j n_j^2}{M} \right\} = \bar{n} - \frac{s^2(n_j)}{N \bar{n}}.$$

For dichotomous outcome variables, the observed between-groups variance is closely related to the chi-squared test statistic (17.6). They are connected by the formula

$$S_{\text{between}}^2 = \frac{\hat{P} (1 - \hat{P})}{\tilde{n} (N - 1)} X^2.$$

The within-groups variance in the dichotomous case is a function of the group averages, namely,

$$S_{\text{within}}^2 = \frac{1}{M-N} \sum_{j=1}^N n_j \bar{Y}_j (1 - \bar{Y}_j).$$

Example 17.1 Religiosity across the world.

This example is about differences in religiosity between countries, as explained by individual- and country-level characteristics. The example follows the paper by Ruiter and van Tubergen (2009), which is about the question why people in some nations attend religious meetings more frequently than people in others. They use data from the European Values Surveys/World Values Surveys collected in 1990–2001 for 136,611 individual respondents in 60 countries.

The dependent variable, *religious attendance*, is whether the respondent attends religious services at least once a week. The number of respondents per country ranges from 95 (Ghana) to 6,501 (Russian Federation). In the data set as a whole, 23.8% of the respondents reported that they attended religious services at least once a week. Ruiter and van Tubergen (2009) formulate various different theories about religiosity and test hypotheses derived from these theories.

A two-level structure is used with the country as the second-level unit. This is based on the idea that there may be differences between countries that are not captured by the explanatory variables and hence may be regarded as unexplained variability within the set of all countries. Figure 17.1 shows for each of the 60 countries the proportion of people in the sample who attend religious services at least once a week. With these large group sizes, a statistical test is not needed to conclude that these proportions, ranging from 0.025 (Denmark) to 0.874 (Nigeria), are very different indeed. The chi-squared test (17.6) for equality of these 60 proportions yields $X^2 = 29,732.52$, $df = 59$, significant at any meaningful significance level.

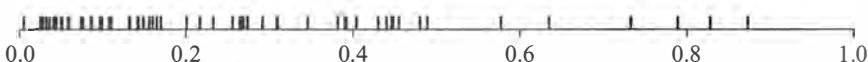


Figure 17.1: Proportion of religious attendance.

The estimated true variance between the country-dependent proportions calculated from (3.11) is $\hat{\tau}^2 = 0.0404$, thus the estimated true between-country standard deviation is $\hat{\tau} = \sqrt{0.0404} = 0.201$. The groups are so large that $\hat{\tau}^2$ is hardly less than S_{between}^2 . With an average probability of 0.238, this standard deviation is quite large and points to considerable skewness of the distribution, as we can see in Figure 17.1.

17.2.2 The logit function: Log-odds

It can be relevant to include explanatory variables in models for dichotomous outcome variables. For example, in the example above, individuals with religious parents will tend to have a higher religious attendance than others. When explanatory variables are included to model probabilities, one problem (mentioned in Section 17.1) is that probabilities are restricted to the domain between 0 and 1, whereas a linear effect for an explanatory variable could take the fitted value outside this interval.

Instead of the probability of some event, one may consider the *odds*: the ratio of the probability of success to the probability of failure. When the probability is p , the odds are

logit(p)

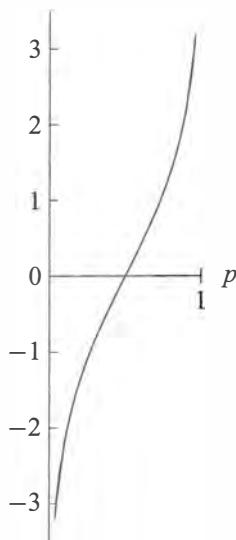


Figure 17.2: The logit function.

$p/(1 - p)$. In contrast to probabilities, odds can assume any value from 0 to infinity, and odds can be considered to constitute a ratio scale.

The *logarithm* transforms a multiplicative to an additive scale and transforms the set of positive real numbers to the whole real line. Indeed, one of the most widely used transformations of probabilities is the *log-odds*, defined by

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right), \quad (17.8)$$

where $\ln(x)$ denotes the natural logarithm of the number x . The logit function, of which a graph is shown in Figure 17.2, is an increasing function defined for numbers between 0 and 1, and its range is from minus infinity to plus infinity. Figure 17.3 shows in a different way in which probability values are transformed to logit values.

For example, $p = 0.269$ is transformed to $\text{logit}(p) = -1$ and $p = 0.982$ to $\text{logit}(p) = 4$. The logit of $p = 0.5$ is exactly 0.

The logistic regression model is a model where $\text{logit}(p)$ is a linear function of the explanatory variables. In spite of the attractive properties of the logit function, it is by no means the only suitable function for transforming probabilities to arbitrary real values. The general term for such a transformation function is the *link function*, as it links the probabilities (or more generally, the expected values of the dependent variable) to the explanatory variables. The probit function (which is the inverse cumulative distribution function of the standard normal distribution) also is often used as a link function for dichotomous variables (see also Section 17.3.2). A generalized linear model for a dichotomous outcome with the probit link function is called a probit regression model. For still other link functions see, for example, Long (1997) or McCullagh and Nelder (1989).

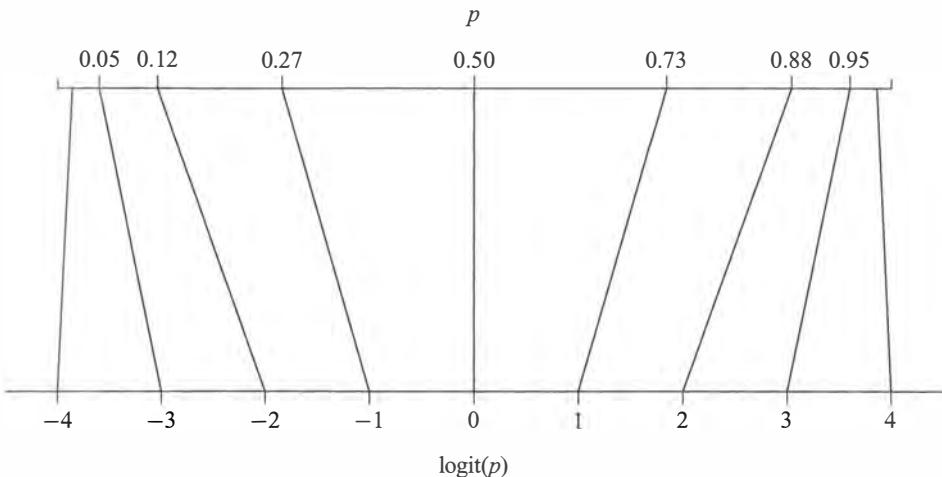


Figure 17.3: Correspondence between p and $\text{logit}(p)$.

The choice of link function has to be guided by the empirical fit of the model, ease of interpretation, and convenience (e.g., availability of computer software). In this chapter the link function for a probability will be denoted by $f(p)$, and we shall concentrate on the logit link function.

17.2.3 The empty model

The empty two-level model for a dichotomous outcome variable refers to a population of groups (level-two units) and specifies the probability distribution for the group-dependent probabilities P_j in (17.2), without taking further explanatory variables into account. Several such specifications have been proposed.

We focus on the model that specifies the transformed probabilities $f(P_j)$ to have a normal distribution. This is expressed, for a general link function $f(p)$, by the formula

$$f(P_j) = \gamma_0 + U_{0j}, \quad (17.9)$$

where γ_0 is the population average of the transformed probabilities and U_{0j} the random deviation from this average for group j . If $f(p)$ is the logit function, then $f(P_j)$ is just the log-odds for group j . Thus, for the logit link function, the log-odds have a normal distribution in the population of groups, which is expressed by

$$\text{logit}(P_j) = \gamma_0 + U_{0j}, \quad (17.10)$$

For the deviations U_{0j} it is assumed that they are independent random variables with a normal distribution with mean 0 and variance τ_0^2 .

This model does not include a separate parameter for the level-one variance. This is because the level-one residual variance of the dichotomous outcome variable follows directly from the success probability, as indicated by equation (17.3).

Denote by π_0 the probability corresponding to the average value γ_0 , as defined by

$$f(\pi_0) = \gamma_0.$$

For the logit function, this means that π_0 is the so-called logistic transform of γ_0 , defined by

$$\pi_0 = \text{logistic}(\gamma_0) = \frac{\exp(\gamma_0)}{1 + \exp(\gamma_0)}. \quad (17.11)$$

Here $\exp(\gamma_0) = e^{\gamma_0}$ denotes the exponential function, where e is the basis of the natural logarithm. The logistic and logit functions are mutual inverses, just like the exponential and the logarithmic functions. Figure 17.4 shows the shape of the logistic function. This π_0 is close (but not quite equal) to the average value of the probabilities P_j in the population of groups. Because of the nonlinear nature of the link function, there is no simple relation between the variance of these probabilities and the variance of the deviations U_{0j} . There is an approximate formula, however, valid when the variances are small. The approximate relation (valid for small τ_0^2) between the population variances is

$$\text{var}(P_j) \approx \frac{\tau_0^2}{(f'(\pi_0))^2}. \quad (17.12)$$

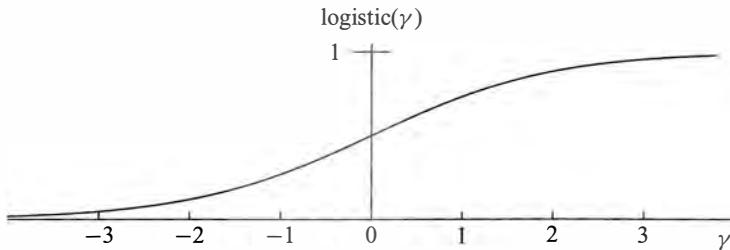


Figure 17.4: The logistic function.

For the logit function, this yields

$$\text{var}(P_j) \approx (\pi_0(1 - \pi_0))^2 \tau_0^2. \quad (17.13)$$

When τ_0^2 is not so small, the variance of the probabilities will be less than the right-hand side of (17.13). (Note that these are population variances and not variances of the observed proportions in the groups; see Section 3.3 for this distinction.)

Example 17.2 Empty model for the religious attendance data.

For the further analyses of the religious attendance data, for one of the countries (Turkey) the income data seemed incorrectly transformed, and therefore this country was omitted from the data set. This leaves 135,508 respondents in 59 countries. Fitting the empty model with normally distributed log-odds to the data of religious attendance in 59 countries yields the results presented in Table 17.1. Note that the table gives the level-two standard deviation, not the variance¹. The model was estimated using the Laplace approximation (see Section 17.2.5).

¹This is because in further modeling of this data set it will be preferable to present standard deviations rather than variances of random slopes.

Table 17.1: Estimates for empty logistic model.

Fixed effect	Coefficient	S.E.
$\gamma_0 = \text{Intercept}$	-1.447	0.180
Random effect	S.D.	
<i>Level-two standard deviation:</i>		
$\tau_0 = \text{S.D.}(U_{0j})$	1.377	
Deviance	120,768.8	

The estimated average log-odds is -1.447 , which corresponds (according to equation (17.11)) to a probability of $\pi_0 = 0.19$, different from the overall fraction of 0.24 because of the nonlinearity of the link function and the weighting of the countries inherent in the multilevel model. The variance approximation formula (17.13) yields $\text{var}(P_j) \approx 0.0451$, not too far from the nonparametric estimate calculated from (3.11), which is 0.0405 . The difference between these values is caused by the facts that formula (17.13) is only an approximation and that the estimation methods are different.

Figure 17.5 presents the observed log-odds of the 59 countries with the fitted normal distribution. These observed log-odds are the logit transformations of the proportions depicted in Figure 17.1. It can be seen that the skewness to the right in Figure 17.1 has more or less disappeared because of the log-odds transformation. There now is one outlier to the left, which diminishes the quality of the normal approximation. However, nonnormality of the level-two residuals is not a serious issue at the stage where we are fitting the empty model, of which we know that it is likely to be misspecified, because no predictor variables are included. The normality assumption will become important mainly when considering residuals in a model with predictor variables.

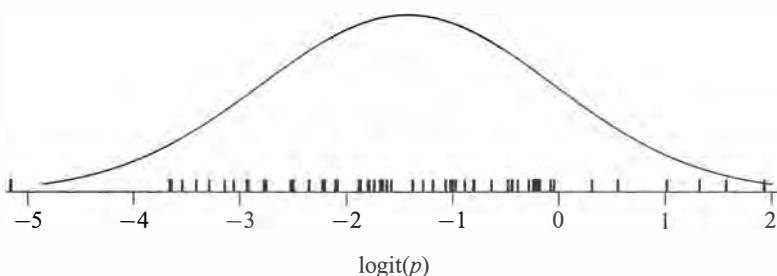


Figure 17.5: Observed log-odds and estimated normal distribution of population log-odds of religious attendance.

17.2.4 The random intercept model

In logistic regression analysis, linear models are constructed for the log-odds. The multilevel analog, random coefficient logistic regression, is based on linear models for the log-odds that include random effects for the groups or other higher-level units. As mentioned above, Y_{ij} denotes the dichotomous outcome variable for level-one unit i in level-two unit j . There are n_j level-one units in the j th level-two unit. The outcome Y_{ij} is coded as

0 or 1, also referred to as ‘failure’ and ‘success’. We shall use the terms ‘individual’ and ‘group’ to refer to the level-one and level-two units.

We now assume that there are variables which are potential explanations for the observed success or failure. These variables are denoted by X_1, \dots, X_r . The values of X_h ($h = 1, \dots, r$) are indicated in the usual way by x_{hij} . Since some (or all) of these variables could be level-one variables, the success probability is not necessarily the same for all individuals in a given group. Therefore the success probability now depends on the individual as well as on the group, and is denoted by P_{ij} . Accordingly, equation (17.2) expressing how the outcome is split into an expected value and a residual now is replaced by

$$Y_{ij} = P_{ij} + R_{ij}. \quad (17.14)$$

The logistic random intercept model expresses the log-odds, that is, the logit of P_{ij} , as a sum of a linear function of the explanatory variables² and a random group-dependent deviation U_{0j} :

$$\text{logit}(P_{ij}) = \gamma_0 + \sum_{h=1}^r \gamma_h x_{hij} + U_{0j}. \quad (17.15)$$

Thus, a unit difference between the X_h -values of two individuals in the same group is associated with a difference of γ_h in their log-odds, or equivalently, a ratio of $\exp(\gamma_h)$ in their odds. The deviations U_{0j} are assumed to have zero means (given the values of all explanatory variables) and variances equal to τ_0^2 . Formula (17.15) does not include a level-one residual because it is an equation for the probability P_{ij} rather than for the outcome Y_{ij} . The level-one residual is already included in formula (17.14).

Example 17.3 Random intercept model for religious attendance.

Continuing the study of religious attendance in 59 countries, we use some of the explanatory variables also considered by Ruiter and van Tubergen (2009). These authors test a number of hypotheses derived from several different theories about religiosity. For these theories and hypotheses, we refer to their paper. Here we use the following variables, some of which have been approximately centered. At the individual level:

- * *Educational level*, measured as the age at which people left school (14–21 years), minus 14. This variable was centered within countries. The within-country deviation variable has mean 0, standard deviation 2.48.
- * *Income*, standardized within country; mean -0.03 , standard deviation 0.99.
- * *Employment status*, 1 for unemployed, 0 for employed; mean 0.19, standard deviation 0.39.
- * *Sex*, 1 for female, 0 for male; mean 0.52, standard deviation 0.50.
- * *Marital status*, 1 for single/divorced/widowed, 0 for married/cohabiting; mean 0.23, standard deviation 0.42.
- * *Divorce status*, 1 for divorced, 0 for other; mean 0.06, standard deviation 0.25.
- * *Widowed*, 1 for widowed, 0 for other; mean 0.08, standard deviation 0.27.
- * *Urbanization*, the logarithm of the number of inhabitants in the community or town of residence, truncated between 1,000 and 1,000,000, minus 10; mean 0.09, standard deviation 2.18.

²Rather than the double-subscript notation γ_{hk} used earlier in Chapters 4 and 5, we now use – to obtain a relatively simple notation – a single-subscript notation γ_h analogous to (5.15).

At the country level, the averages of some individual-level variables representing social and cultural differences between countries were used, with one other country-level variable. Using the definitions above, this leads to the following list of variables:

- * Average educational level: mean 0.82, standard deviation 0.94.
- * Average unemployment: mean 0.19, standard deviation 0.08.
- * Average divorce status: mean 0.06, standard deviation 0.03.
- * *Gini coefficient* measuring income inequality, minus 35.
Mean 0.10, standard deviation 9.58.

Parameters were estimated by the Laplace algorithm implemented in R package lme4 (see Section 17.2.5). The results are shown in Table 17.2.

Table 17.2: Logistic random intercept model for religious attendance in 59 countries.

Fixed effect	Model 1	
	Coefficient	S.E.
γ_0 Intercept	-2.069	0.646
γ_1 Education (within-country deviation)	-0.0290	0.0032
γ_2 Income	-0.0638	0.0082
γ_3 Unemployed	0.017	0.020
γ_4 Female	0.508	0.016
γ_5 Single	-0.269	0.019
γ_6 Divorced	-0.489	0.036
γ_7 Widowed	0.518	0.027
γ_8 Urbanization	-0.0665	0.0039
γ_9 Gini coefficient	0.035	0.017
γ_{10} Country average education	-0.330	0.135
γ_{11} Country average unemployment	6.033	2.116
γ_{12} Country average divorce	-7.120	4.975
Random effect	S.D.	S.E.
<i>Random intercept:</i>		
$\tau_0 = \text{S.D.}(\bar{U}_{0j})$ intercept standard deviation	1.08	
Deviance	115,969.9	

All the individual-level variables have significant effects, except for unemployment, which only has an effect at the country level. For education and divorce, both of which have negative effects on religious attendance, the between-country regression coefficients are stronger than the within-country coefficients. The fact that regression coefficients of the country average variables are larger in size than the coefficients of individual variables must be seen in the light that these are unstandardized coefficients, and the country averages have considerably smaller variability than the individual-level variables.

17.2.5 Estimation

Parameter estimation in hierarchical generalized linear models is more complicated than in hierarchical linear models. Inevitably some kind of approximation is involved, and various kinds of approximation have been proposed. Good reviews are given in Demidenko (2004, Chapter 8), Skrondal and Rabe-Hesketh (2004, Chapter 6), Tuerlinckx et al. (2006), and Rodríguez (2008). We mention some references and some of the terms used without explaining them. The reader who wishes to study these algorithms is referred to the literature cited. More information about the computer programs mentioned in this section is given in Chapter 18.

Currently the best and most often used methods are two frequentist ones: the Laplace approximation and adaptive numerical quadrature as algorithms for approximating the maximum likelihood estimates; and Bayesian methods.

The *Laplace approximation* for hierarchical generalized linear models was proposed by Raudenbush et al. (2000). Because of its good quality and computational efficiency this is now one of the most frequently used estimation algorithms for generalized linear mixed models. It is implemented in the software package HLM and in various packages of the statistical system R, such as lme4, glmmADMB and glmmML.

Numerical integration is an approach for dealing with the random coefficients which is straightforward in principle, but poses various difficulties in implementation. Its use for hierarchical generalized linear models was initially developed by Stiratelli et al. (1984), Anderson and Aitkin (1985) and Gibbons and Bock (1987); later publications include Longford (1994), Hedeker and Gibbons (1994), Gibbons and Hedeker (1997), and Hedeker (2003). Pinheiro and Bates (1995) and Rabe-Hesketh et al. (2005) developed so-called *adaptive quadrature* methods of numerical integration; see also Skrondal and Rabe-Hesketh (2004). Adaptive quadrature has considerable advantages over nonadaptive numerical integration. It is implemented in SAS proc NLMIXED, the Stata package, and the software gllamm which combines with Stata. Nonadaptive numerical integration is implemented in the program MIXOR/SuperMix and its relatives.

Much work has been done on Bayesian methods (cf. Section 12.1), mainly used in Markov chain Monte Carlo (MCMC) methods, which are computationally intensive, because they are based on repeated simulation of draws from the posterior distribution. Zeger and Karim (1991) and Browne and Draper (2000) are important milestones for Bayesian inference for multilevel models, and this approach is now being increasingly used. An overview is given by Draper (2008) and an extensive treatment in Congdon (2010). For hierarchical generalized linear models, Bayesian methods perform very well, and they also have good frequentist properties (i.e., small mean squared errors of estimators and good coverage probabilities of confidence intervals); see Browne and Draper (2006). Bayesian MCMC methods are implemented in MLwiN, WinBUGS, and BayesX. A recent development is the use of the Laplace approximation for Bayesian inference, which circumvents the computationally intensive MCMC methods; see Rue et al. (2009) and Fong et al. (2010). This is implemented in the R package INLA.

Before these methods had been developed, the main methods available were those based on first- or second-order Taylor expansions of the link function. When the approximation is around the estimated fixed part, this is called marginal quasi-likelihood (MQL), when it is around an estimate for the fixed plus the random part it is called penalized or predictive quasi-likelihood (PQL) (Breslow and Clayton, 1993; Goldstein, 1991; Goldstein and

Rasbash, 1996). For estimating and testing fixed effects these methods are quite adequate, especially if the cluster sizes n_j are not too small, but they are not satisfactory for inference about random effects. The first-order MQL and PQL estimates of the variance parameters of the random part have an appreciable downward bias (Rodríguez and Goldman, 1995; Browne and Draper, 2006; Rodríguez, 2008; Austin, 2010). The second-order MQL and PQL methods produce parameter estimates with less bias but, it seems, a higher mean squared error. The biases of MQL and PQL in the estimation of parameters of the random effects can be diminished by bootstrapping (Kuk, 1995; van der Leeden et al., 2008), but this leads to quite computer-intensive procedures. These methods are implemented in MLwiN, HLM, R packages MASS and nlme, and SAS proc GLIMMIX.

Several other methods have been proposed for parameter estimation in hierarchical generalized linear models. McCulloch (1997), Ng et al. (2006), and Jank (2006) gave various algorithms for estimation by simulated maximum likelihood. Another computer-intensive method is the method of simulated moments. This method is applied to these models by Gouriéroux and Montfort (1996, Section 3.1.4 and Chapter 5), and an overview of some recent work is given by Baltagi (2008, Section 11.2). A method based on the principle of indirect inference was proposed by Mealli and Rampichini (1999).

Various simulation studies have been done comparing different estimation procedures for hierarchical generalized linear models, mostly focusing on multilevel logistic regression: Rodriguez and Goldman (1995, 2001), Browne and Draper, (2006), Rodríguez, (2008), and Austin (2010). These are the references on which the statements in this section about the qualities of the various estimation procedures are based.

17.2.6 Aggregation

If the explanatory variables assume only few values, then it is advisable to aggregate the individual 0–1 data to success counts, depending on the explanatory variables, within the level-two units. This will improve the speed and stability of the algorithm and reduce memory use. This is carried out as follows.

For a random intercept model with a small number of discrete explanatory variables X_1, \dots, X_r , let L be the total number of combinations of values (x_1, \dots, x_r) . All individuals with the same combination of values (x_1, \dots, x_r) are treated as one subgroup in the data. They all have a common success probability, given by (17.15). Thus, each level-two unit includes L subgroups, or fewer if some of the combinations do not occur in this level-two unit. Aggregation is advantageous if L is considerably less than the average group size n_j .

Denote by

$$n_j^+(x_1, \dots, x_r)$$

the number of individuals in group j with the values x_1, \dots, x_r on the respective explanatory variables, and denote by

$$Y_j^+(x_1, \dots, x_r)$$

the number of individuals among these who yielded a success, that is, for whom $Y_{ij} = 1$. Then $Y_j^+(x_1, \dots, x_r)$ has the binomial distribution with binomial denominator ('number of trials') $n_j^+(x_1, \dots, x_r)$ and success probability given by (17.15), which is the same for all individuals i in this subgroup. The multilevel analysis is now applied with these subgroups as the level-one units. Subgroups with $n_j^+(x_1, \dots, x_r) = 0$ can be omitted from the data set.

17.3 Further topics on multilevel logistic regression

17.3.1 Random slope model

The random intercept logistic regression model of Section 17.2.4 can be extended to a random slope model just as in Chapter 5. We only give the formula for one random slope; the extension to several random slopes is straightforward. The remarks made in Chapters 5 and 6 remain valid, given the appropriate changes, and will not be repeated here.

As in the random intercept model, assume that there are r explanatory variables X_1, \dots, X_r . Assume that the effect of X_1 is variable across groups, and accordingly has a random slope. Expression (17.15) for the logit of the success probability is extended with the random effect $U_{1j}x_{1j}$, which leads to

$$\text{logit}(P_{ij}) = \gamma_0 + \sum_{h=1}^r \gamma_h x_{hij} + U_{0j} + U_{1j}x_{1ij}. \quad (17.16)$$

There now are two random group effects, the random intercept U_{0j} and the random slope U_{1j} . It is assumed that both have a zero mean. Their variances are denoted, respectively, by τ_0^2 and τ_1^2 and their covariance is denoted by τ_{01} .

Example 17.4 Random slope model for religious attendance.

Continuing the example of religious attendance in 59 countries, Table 17.3 presents a model which has the same fixed part as the model of Table 17.2, and random slopes for income and education. Recall that income is standardized within countries and education is a within-country deviation variable. Estimating a model with a random slope for education was successful for this relative measure of education, but not when the grand-mean-centered education variable was used; this may be a signal of misspecification of the latter model.

The deviance goes down by $115,969.9 - 115,645.7 = 324.2$, a huge improvement for five extra parameters. Also when each of the random slope variances is considered separately, the gain in deviance is considerable (results not shown here). With such large sample sizes per country, the regression coefficients within each country are estimated very precisely and it is not surprising to find evidence for slope heterogeneity.

For education as well as income, the slope standard deviation is larger in absolute magnitude than the estimated fixed effect, which is negative for both variables. This implies that for most countries the regression coefficient per country is negative, but for a nonnegligible number of countries it is positive. For example, the regression coefficient of income is $\gamma_2 + U_{2j}$, which has an estimated normal distribution with mean -0.074 and standard deviation 0.063 , which leads to a probability of 0.12 of a randomly chosen country having a positive coefficient.

The standard errors of the fixed effects of education and income are considerably larger here than in the random intercept model of Table 17.2. The random slope model must be considered to be more realistic, and therefore the larger standard errors are a better representation of the uncertainty about these fixed parameters, which are the estimated average effects in the population of all countries. This suggests that for other level-one variables a random slope could, or should, also be considered, and their estimated coefficients might then also get higher standard errors. For a data set with very large level-two units such as this one, the two-step approach based on country-by-country equations of the type (3.38)–(3.40) – but of course with a larger number of predictor variables – may be more suitable than the approach by a hierarchical generalized linear model, because less stringent assumptions are being made about the similarity of many parameters across the various different countries (cf. Achen, 2005).

Table 17.3: Logistic random intercept model for religious attendance
in 59 countries.

Fixed effect	Model 1	
	Coefficient	S.E.
γ_0 Intercept	3.792	2.476
γ_1 Education (within-country deviation)	-0.0388	0.0092
γ_2 Income	-0.0738	0.0161
γ_3 Unemployed	0.019	0.020
γ_4 Female	0.511	0.016
γ_5 Single	-0.271	0.019
γ_6 Divorced	-0.493	0.036
γ_7 Widowed	0.482	0.027
γ_8 Urbanization	-0.0650	0.0040
γ_9 Gini coefficient	0.028	0.017
γ_{10} Country average education	-0.333	0.132
γ_{11} Country average unemployment	5.44	2.06
γ_{12} Country average divorce	-6.43	4.84
Random part parameters	S.D. / Corr.	
$\tau_0 = S.D.(U_{0j})$ Intercept standard deviation	1.09	
$\tau_1 = S.D.(U_{1j})$ Income-slope standard deviation	0.096	
$\tau_0 = S.D.(U_{2j})$ Education-slope standard deviation	0.063	
$\rho_{01} = \rho(U_{0j}, U_{1j})$ Intercept-income slope correlation	0.29	
$\rho_{02} = \rho(U_{0j}, U_{2j})$ Intercept-education slope correlation	-0.07	
$\rho_{12} = \rho(U_{1j}, U_{2j})$ Income-education slopes correlation	0.27	
Deviance	115,645.7	

17.3.2 Representation as a threshold model

The multilevel logistic regression can also be formulated as a so-called *threshold model*. The dichotomous outcome Y , ‘success’ or ‘failure’, is then conceived as the result of an underlying non-observed continuous variable. When Y denotes passing or failing some test or exam, the underlying continuous variable could be the scholastic aptitude of the subject; when Y denotes whether the subject behaves in a certain way, the underlying variable could be a variable representing benefits minus costs of this behavior; etc. Denote the underlying variable by \tilde{Y} . Then the threshold model states that Y is 1 if \tilde{Y} is larger than some threshold, and 0 if it is less than the threshold. Since the model is about unobserved entities, it is no restriction to assume that the threshold is 0. This leads to the representation

$$Y = \begin{cases} 1 & \text{if } \tilde{Y} > 0 \\ 0 & \text{if } \tilde{Y} \leq 0. \end{cases} \quad (17.17)$$

For the unobserved variable \check{Y} , a usual random intercept model is assumed:

$$\check{Y}_{ij} = \gamma_0 + \sum_{h=1}^r \gamma_h x_{hij} + U_{0j} + R_{ij}. \quad (17.18)$$

To represent a logistic regression model, the level-one residual of the underlying variable \check{Y} must have a *logistic distribution*. This means that, when the level-one residual is denoted by R_{ij} , the cumulative distribution function of R_{ij} must be the logistic function,

$$P(R_{ij} < x) = \text{logistic}(x), \quad \text{for all } x, \quad (17.19)$$

defined in (17.11). This is a symmetric probability distribution, so that also

$$P(-R_{ij} < x) = \text{logistic}(x), \quad \text{for all } x.$$

Its mean is 0 and its variance is $\pi^2/3 = 3.29$. When it is assumed that R_{ij} has this distribution, the logistic random intercept model (17.15) is equivalent to the threshold model defined by (17.17) and (17.18).

To represent the random slope model (17.16) as a threshold model, we define

$$\check{Y}_{ij} = \gamma_0 + \sum_{h=1}^r \gamma_h x_{hij} + U_{0j} + U_{1j} x_{1ij} + R_{ij}, \quad (17.20)$$

where R_{ij} has a logistic distribution. It then follows that

$$\begin{aligned} P(Y_{ij} = 1) &= P(\check{Y}_{ij} > 0) \\ &= P\left(-R_{ij} < \gamma_0 + \sum_{h=1}^r \gamma_h x_{hij} + U_{0j} + U_{1j} x_{1ij}\right) \\ &= \text{logistic}\left(\gamma_0 + \sum_{h=1}^r \gamma_h x_{hij} + U_{0j} + U_{1j} x_{1ij}\right). \end{aligned}$$

Since the logit and the logistic functions are mutual inverses, the last equation is equivalent to (17.16).

If the residual R_{ij} has a standard normal distribution with unit variance, then the probit link function is obtained. Thus, the threshold model which specifies that the underlying variable \check{Y} has a distribution according to the hierarchical linear model of Chapters 4 and 5, with a normally distributed level-one residual, corresponds exactly to the multilevel probit regression model. Since the standard deviation of R_{ij} is $\sqrt{\pi^2/3} = 1.81$ for the logistic and 1 for the probit model, the fixed estimates for the logistic model will tend to be about 1.81 times as large as for the probit model and the variance parameters of the random part about $\pi^2/3 = 3.29$ times as large (but see Long (1997, p. 48), who notes that in practice the proportionality constant for the fixed estimates is closer to 1.7).

17.3.3 Residual intraclass correlation coefficient

The intraclass correlation coefficient for the multilevel logistic model can be defined in at least two ways. The first definition is by applying the definition in Section 3.3 straightforwardly to the binary outcome variable Y_{ij} . This approach was also mentioned in Section 17.2.1. It was followed, for example, by Commenges and Jacqmin (1994).

The second definition is by applying the definition in Section 3.3 to the unobserved underlying variable \check{Y}_{ij} . Since the logistic distribution for the level-one residual implies a variance of $\pi^2/3 = 3.29$, this implies that for a two-level logistic random intercept model with an intercept variance of τ_0^2 , the intraclass correlation is

$$\rho_I = \frac{\tau_0^2}{\tau_0^2 + \pi^2/3}.$$

These two definitions are different and will lead to somewhat different outcomes. For example, for the empty model for the religious attendance data presented in Table 17.1, the first definition yields $0.0404/(0.0404+0.1414) = 0.22$, whereas the second definition leads to the value $1.896/(1.896+3.290) = 0.37$. In this case, the difference is large, underscoring the rather arbitrary nature of the definition of these coefficients.

An advantage of the second definition is that it can be directly extended to define the residual intraclass correlation coefficient, that is, the intraclass correlation which controls for the effect of explanatory variables. The example can be continued by moving to the model in Table 17.2. The residual intraclass correlation controlling for the variables in this model is $1.177/(1.177+3.290) = 0.26$, lower than the raw intraclass correlation coefficient.

For the multilevel probit model, the second definition for the intraclass correlation (and its residual version) leads to

$$\rho_I = \frac{\tau_0^2}{\tau_0^2 + 1},$$

since this model fixes the level-one residual variance of the unobservable variable \check{Y}_{ij} to 1.

17.3.4 Explained variance

There are several definitions of the explained proportion of variance (R^2) in single-level logistic and probit regression models. Reviews are given by Hagle and Mitchell (1992), Veall and Zimmermann (1992), and Windmeijer (1995). Long (1997, Section 4.3) presents an extensive overview. One of these definitions, the R^2 measure of McKelvey and Zavoina (1975), which is based on the threshold representation treated in Section 17.3.2, is considered very attractive in each of these reviews. In this section we propose a measure for the explained proportion of variance which extends McKelvey and Zavoina's measure to the logistic and probit random intercept model.

It is assumed that it makes sense to conceive of the dichotomous outcomes Y as being generated through a threshold model with underlying variable \check{Y} . In addition, it is assumed that the explanatory variables X_h are random variables. (This assumption is always made for defining the explained proportion of variance; see the introduction of Section 7.1.) Therefore the explanatory variables are, as in Chapter 7, indicated by capital letters. For the underlying variable \check{Y} , equation (17.18) gives the expression

$$\check{Y}_{ij} = \gamma_0 + \sum_{h=1}^r \gamma_h X_{hij} + U_{0j} + R_{ij}.$$

Denote the fixed part by

$$\hat{Y}_{ij} = \gamma_0 + \sum_{h=1}^r \gamma_h X_{hij}. \quad (17.21)$$

This variable is also called the *linear predictor* for Y . Its variance is denoted by σ_F^2 . The intercept variance is $\text{var}(U_{0j}) = \tau_0^2$ and the level-one residual variance is denoted by $\text{var}(R_{ij}) = \sigma_R^2$. Recall that σ_R^2 is fixed to $\pi^2/3 = 3.29$ for the logistic and to 1 for the probit model.

For a randomly drawn level-one unit i in a randomly drawn level-two unit j , the X -values are randomly drawn from the corresponding population and hence the total variance of \check{Y}_{ij} is equal to

$$\text{var}(\check{Y}_{ij}) = \sigma_F^2 + \tau_0^2 + \sigma_R^2.$$

The explained part of this variance is σ_F^2 and the unexplained part is $\tau_0^2 + \sigma_R^2$. Of this unexplained variation, τ_0^2 resides at level two and σ_R^2 at level one. Hence the proportion of explained variation can be defined by

$$R_{\text{dicho}}^2 = \frac{\sigma_F^2}{\sigma_F^2 + \tau_0^2 + \sigma_R^2}. \quad (17.22)$$

The corresponding definition of the residual intraclass correlation,

$$\rho_I = \frac{\tau_0^2}{\tau_0^2 + \sigma_R^2}, \quad (17.23)$$

was also given in Section 17.3.3.

Estimating (17.22) is done in three steps. First one computes, as a new transformed variable, the linear predictor \hat{Y}_{ij} defined in (17.21) using the estimated coefficients $\hat{\gamma}_0, \hat{\gamma}_1, \dots, \hat{\gamma}_r$. Next the variance σ_F^2 is estimated by computing the observed variance of this new variable. Finally, this value is plugged into (17.22) together with the estimated intercept variance $\hat{\tau}_0^2$ and the fixed value of σ_R^2 . The example below gives an illustration.

In the interpretation of this R^2 -value it should be kept in mind that such values are known for single-level logistic regression to be usually considerably lower than the OLS R^2 values obtained for predicting continuous outcomes.

Example 17.5 Explained variance for religious attendance.

For the random intercept model in Table 17.2, the linear predictor is

$$\begin{aligned} \hat{Y}_{ij} = & -2.069 - 0.029X_{1ij} - 0.0638X_{2ij} + 0.017X_{3ij} + 0.508X_{4ij} \\ & - 0.269X_{5ij} - 0.489X_{6ij} + 0.518X_{7ij} - 0.0665X_{8ij} \\ & + 0.035X_{9j} - 0.33X_{10,ij} + 6.033X_{11,ij} - 7.12X_{12,ij}, \end{aligned}$$

using the symbols X_1, \dots, X_{12} for the predictor variables. This constructed variable has variance $\hat{\sigma}_F^2 = 1.049$ (a value obtained by calculating this new variable and computing its variance). Equation (17.22) yields the proportion of explained variance $R_{\text{dicho}}^2 = 1.049/(1.049 + 1.08^2 + 3.29) = 0.19$.

Example 17.6 Taking a science subject in high school.

This example continues the analysis of the data set of Example 8.3 about the cohort of pupils entering secondary school in 1989, studied by Dekkers et al. (2000). The focus now is on whether the pupils chose at least one science subject for their final examination. The sample is restricted to pupils in general education (excluding junior vocational education), and to only those who progressed to their final examination (excluding drop-outs and pupils who repeated grades once or twice). This left 3,432

pupils distributed over 240 secondary schools. There were 736 pupils who took no science subjects, 2,696 who took one or more.

A multilevel logistic regression model was estimated by R package lme4 and the MIXOR program. Both gave almost the same results; the MIXOR results are presented. Explanatory variables are gender (0 for boys, 1 for girls) and minority status (0 for children of parents born in industrialized countries, 1 for other countries). The results are shown in Table 17.4.

Table 17.4: Estimates for probability of taking a science subject.

Fixed effect	Model 1	
	Coefficient	S.E.
γ_0 Intercept	2.487	0.110
γ_1 Gender	-1.515	0.102
γ_2 Minority status	-0.727	0.195
Random effect	Var. comp.	S.E.
<i>Level-two variance:</i>		
$\tau_0^2 = \text{var}(U_{0j})$	0.481	0.082
Deviance	3,238.27	

The linear predictor for this model is

$$\hat{Y}_{ij} = 2.487 - 1.515 \text{Gender}_{ij} - 0.727 \text{Minority}_{ij}$$

and the variance of this variable in the sample is $\hat{\sigma}_F^2 = 0.582$. Therefore the explained proportion of variation is

$$R^2_{\text{dicho}} = \frac{0.582}{0.582 + 0.481 + 3.29} = 0.13.$$

In other words, gender and minority status explain about 13% of the variation in whether the pupil takes at least one science subject for the high school exam.

The unexplained proportion of variation, $1 - 0.13 = 0.87$, can be written as

$$\frac{0.481}{0.582 + 0.481 + 3.29} + \frac{3.29}{0.582 + 0.481 + 3.29} = 0.11 + 0.76 = 0.87,$$

which represents the fact that 11% of the variation is unexplained variation at the school level and 76% is unexplained variation at the pupil level. The residual intraclass correlation is $\rho_I = 0.481 / (0.481 + 3.29) = 0.13$.

17.3.5 Consequences of adding effects to the model

When a random intercept is added to a given logistic or probit regression model, and also when variables with fixed effects are added to such a model, the effects of earlier included variables may change. The nature of this change, however, may be different from such changes in OLS or multilevel linear regression models for continuous variables.

This phenomenon can be illustrated by continuing the example of the preceding section. Table 17.5 presents three other models for the same data, in all of which some elements were omitted from Model 1 as presented in Table 17.4.

Table 17.5: Three models for taking a science subject.

Fixed effect	Model 2		Model 3		Model 4	
	Par.	S.E.	Par.	S.E.	Par.	S.E.
γ_0 Intercept	2.246	0.090	2.440	0.109	1.448	0.058
γ_1 Gender	-1.397	0.102	-1.507	0.102		
γ_2 Minority status					-0.644	0.174
Random effect			Var. comp.	S.E.	Var. comp.	S.E.
<i>Level-two variance:</i>						
$\tau_0^2 = \text{var}(U_{0j})$			0.514	0.084	0.293	0.043
Deviance	3,345.15		3,251.86		3,476.06	

Models 2 and 3 include only the fixed effect of gender; Model 2 does not contain a random intercept and therefore is a single-level logistic regression model, while Model 3 does include the random intercept. The deviance difference ($\chi^2 = 3,345.15 - 3,251.86 = 93.29$, $df = 1$) indicates that the random intercept variance is significantly positive. But the sizes of the fixed effects increase in absolute value when adding the random intercept to the model, both by about 8%. Gender is evenly distributed across the 240 schools, and one may wonder why the absolute size of the effect of gender increases when the random school effect is added to the model.

Model 4 differs from Model 1 in that the effect of gender is excluded. The fixed effect of minority status in Model 4 is -0.644 , whereas in Model 1 it is -0.727 . The intercept variance in Model 4 is 0.293 and in Model 1 it is 0.481. Again, gender is evenly distributed across schools and across the majority and minority pupils, and the question is how to interpret the fact that the intercept variance, that is, the unexplained between-school variation, rises, and that also effect of minority status becomes larger in absolute value, when the effect of gender is added to the model.

The explanation can be given on the basis of the threshold representation. If all fixed effects γ_h and also the random intercept U_{0j} and the level-one residual R_{ij} were multiplied by the same positive constant c , then the unobserved variable \check{Y}_{ij} would also be multiplied by c . This corresponds also to multiplying the variances τ_0^2 and σ_R^2 by c^2 . However, it follows from (17.17) that the observed outcome Y_{ij} would not be affected because when \check{Y}_{ij} is positive then so is $c\check{Y}_{ij}$. This shows that the regression parameters and the random part parameters of the multilevel logistic and probit models are meaningful only because the level-one residual variance σ_R^2 has been fixed to some value; but this value is more or less arbitrary because it is chosen merely by the convention of $\sigma_R^2 = \pi^2/3$ for the logistic and $\sigma_R^2 = 1$ for the probit model.

The meaningful parameters in these models are the *ratios* between the regression parameters γ_h , the random effect standard deviations τ_0 (and possibly τ_1 , etc.), and the level-one residual standard deviation σ_R . Armed with this knowledge, we can understand the consequences of adding a random intercept or a fixed effect to a logistic or probit regression model.

When a single-level logistic or probit regression model has been estimated, the random variation of the unobserved variable \tilde{Y} in the threshold model is σ_R^2 . When subsequently a random intercept is added, this random variation becomes $\sigma_R^2 + \tau_0^2$. For explanatory variables that are evenly distributed between the level-two units, the ratio of the regression coefficients to the standard deviation of the (unexplained) random variation will remain approximately constant. This means that the regression coefficients will be multiplied by about the factor

$$\sqrt{1 + \frac{\tau_0^2}{\sigma_R^2}}.$$

In the comparison between Models 2 and 3 above, this factor is $\sqrt{1+(0.514/3.29)} = 1.08$. This is indeed approximately the number by which the regression coefficients were multiplied, when going from Model 2 to Model 3.

It can be concluded that, compared to single-level logistic or probit regression analysis, including random intercepts tends to increase (in absolute value) the regression coefficients. For single-level models for binary variables, this was discussed by Winship and Mare (1984). In the biostatistical literature, this is known as the phenomenon where population-averaged effects (i.e., effects in models without random effects) are closer to zero than cluster-specific effects (which are the effects in models with random effects). Further discussions can be found in Neuhaus et al. (1991), Neuhaus (1992), Diggle et al. (2002, Section 7.4), and Skrondal and Rabe-Hesketh (2004, Section 4.8). Bauer (2009) proposes a way to put the estimates from different models on a common scale.

Now suppose that a multilevel logistic or probit regression model has been estimated, and the fixed effect of some level-one variable X_{r+1} is added to the model. One might think that this would lead to a decrease in the level-one residual variance σ_R^2 . However, this is impossible as this residual variance is fixed, so that instead the estimates of the other regression coefficients will tend to become larger in absolute value and the intercept variance (and slope variances, if any) will also tend to become larger. If the level-one variable X_{r+1} is uncorrelated with the other included fixed effects and also is evenly distributed across the level-two units (i.e., the intraclass correlation of X_{r+1} is about nil), then the regression coefficients γ_h and the standard deviations τ_0 (etc.) of the random effects will all increase by about the same factor. Correlations between X_{r+1} and other variables or positive intraclass correlation of X_{r+1} may distort this pattern to a greater or lesser extent.

This explains why the effect of minority status and the intercept variance increase when going from Model 4 to Model 1. The standard deviation of the random intercept increases by a larger factor than the regression coefficient of minority status, however. This might be related to an interaction between the effects of gender and minority status and to a very even distribution of the sexes across schools (cf. Section 7.1).

17.4 Ordered categorical variables

Variables that have as outcomes a small number of ordered categories are quite common in the social and biomedical sciences. Examples of such variables are outcomes of questionnaire items (with outcomes such as ‘completely disagree’, ‘disagree’, ‘agree’, ‘completely agree’), a test scored by a teacher as ‘fail’, ‘satisfactory’, or ‘good’, etc. This section is about multilevel models where the dependent variable is such an ordinal variable.

When the number of categories is two, the dependent variable is dichotomous and Section 17.2 applies. When the number of categories is rather large (5 or more), it may be possible to approximate the distribution of the residuals by a normal distribution and apply the hierarchical linear model for continuous outcomes. The main issue in such a case is the homoscedasticity assumption: is it reasonable to assume that the variances of the random terms in the hierarchical linear model are constant? (The random terms in a random intercept model are the level-one residuals and the random intercept, R_{ij} and U_0 , in (4.8).) To check this, it is useful to investigate the skewness of the distribution. If in some groups, or for some values of the explanatory variables, the dependent variable assumes outcomes that are very skewed toward the lower or upper end of the scale, then the homoscedasticity assumption is likely to be violated.

If the number of categories is small (3 or 4), or if it is between 5 and, say, 10, and the distribution cannot well be approximated by a normal distribution, then statistical methods for ordered categorical outcomes can be useful. For single-level data such methods are treated, for example, in McCullagh and Nelder (1989) and Long (1997).

It is usual to assign numerical values to the ordered categories, taking into account that the values are arbitrary. To have a notation that is compatible with the dichotomous case of Section 17.2, the values for the ordered categories are defined as $0, 1, \dots, c - 1$, where c is the number of categories. Thus, in the four-point scale mentioned above, ‘completely disagree’ would get the value 0, ‘disagree’ would be represented by 1, ‘agree’ by 2, and ‘completely agree’ by the value 3. The dependent variable for level-one unit i in level-two unit j is again denoted Y_{ij} , so that Y_{ij} now assumes values in the set $\{0, 1, \dots, c - 1\}$.

A very useful model for this type of data is the *multilevel ordered logistic regression model*, also called the *multilevel ordered logit model* or the *multilevel proportional odds model*; and the closely related *multilevel ordered probit model*. These models are discussed, for example, by Agresti and Natarajan (2001), Gibbons and Hedeker (1994), Hedeker (2008), Hedeker and Gibbons (2006, Chapter 10), Rabe-Hesketh and Skrondal (2008, Chapter 7), and Goldstein (2011). A three-level model was discussed by Gibbons and Hedeker (1997).

These models can be formulated as threshold models as in Section 17.3.2, now with $c - 1$ thresholds rather than one. The real line is divided by the thresholds into c intervals (of which the first and the last have infinite length), corresponding to the c ordered categories. The first threshold is $\theta_0 = 0$, the higher thresholds are denoted $\theta_1, \theta_2, \dots, \theta_{c-2}$. Threshold θ_k defines the boundary between the intervals corresponding to observed outcomes k and $k + 1$ (for $k = 0, 1, \dots, c - 2$). The assumed unobserved underlying continuous variable is again denoted by \check{Y} and the observed categorical variable Y is related to \check{Y} by the ‘measurement model’ defined as

$$Y = \begin{cases} 0 & \text{if } \check{Y} \leq \theta_0 \\ 1 & \text{if } \theta_0 < \check{Y} \leq \theta_1, \\ k & \text{if } \theta_{k-1} < \check{Y} \leq \theta_k \ (k = 2, \dots, c-2), \\ c-1 & \text{if } \theta_{c-2} < \check{Y}. \end{cases} \quad (17.24)$$

For $c = 2$ categories this reduces to just the two-category threshold representation in (17.17).

The random intercept ordered category model with explanatory variables X_1, \dots, X_r is based on the ‘structural model’ for the unobserved underlying variable,

$$\check{Y}_{ij} = \gamma_0 + \sum_{h=1}^r \gamma_h x_{hij} + U_{0j} + R_{ij}. \quad (17.25)$$

The structural model and the measurement model (17.24) together determine the distribution of Y_{ij} . If R_{ij} has the logistic distribution (17.19) this results in the *multilevel ordered logistic regression model*, also called the multilevel ordered logit model or multilevel proportional odds model. If R_{ij} has the standard normal distribution, this leads to the *multilevel ordered probit model*. The differences between these two models are minor and the choice between them is a matter of fit and convenience.

The parameters of the structural model can be interpreted in principle just as in the hierarchical linear model. The intraclass correlation coefficient can be defined as in Section 17.3.3. This definition referred only to the model for the underlying variable \check{Y} , and therefore can be applied immediately to the multi-category case. Similarly, the proportion of explained variance can be defined, as in Section 17.3.4, by the formula

$$R_{\text{poly}}^2 = \frac{\sigma_F^2}{\sigma_F^2 + \tau_0^2 + \sigma_R^2}, \quad (17.26)$$

where σ_F^2 is the variance of the fixed part (or the linear predictor) while σ_R^2 is $\pi^2/3 = 3.29$ for the logistic model and 1 for the probit model.

The threshold parameters are usually of secondary importance and reflect the marginal probabilities of the outcome categories: if category k has a low probability then θ_{k-1} will be not much less than θ_k . For more discussion about the interpretation of the fixed parameters we refer to the literature on the single-level version of this model, such as Long (1997).

The model can be extended with a random slope in a straightforward manner. However, estimation algorithms for these models are less stable than for the standard hierarchical linear model, and it is not uncommon that it is impossible to obtain converging parameter estimates for models with even only one random slope.

What was said in Section 17.3.5 about the effect of adding level-one variables to a multilevel logistic regression model is valid also for the multilevel ordered logit and probit models. When some model has been fitted and an important level-one variable is added to this model, this will tend to increase the level-two variance parameters (especially if the newly added variable explains mainly within-group variation), the threshold parameters, and the absolute sizes of the regression coefficients (especially for variables that are uncorrelated with the newly added variable). This is also discussed in Fielding (2004b) and Bauer (2009).

These models can be estimated by the methods discussed above in Section 17.2.5. Various of these procedures are implemented in the programs MLwiN, HLM, MIXOR, Stata,

and SAS (see Chapter 18). Some computer programs do not put the first threshold equal to 0, but the intercept. Thus, $\gamma_0 = 0$ and this parameter is not estimated, but instead θ_0 is not equal to 0 and is estimated. This is a reparametrization of the model and yields parameters which can simply be translated into one another, since it follows from (17.24) and (17.25) that subtracting the same number of all thresholds as well as of γ_0 yields the same distribution of the observed variables Y_{ij} .

Dichotomization of ordered categories

Models for ordered categorical outcomes are more complicated to fit and to interpret than models for dichotomous outcomes. Therefore it may make sense also to analyze the data after dichotomizing the outcome variable. For example, if there are three outcomes, one could analyze the dichotomization 1 versus {2, 3} and also {1, 2} versus 3. Each of these analyses separately is of course based on less information, but may be easier to carry out and to interpret.

Suppose that the multilevel c -category logistic or probit model, defined by (17.24) and (17.25), is indeed valid. Then for each of the dichotomizations, the population parameters of the structural model (17.25), except for the fixed intercept γ_0 , are also the population parameters of the multilevel logistic regression model. (The fixed intercept in the analysis of the dichotomized outcomes depends on the fixed intercept for the multicategory outcomes together with the threshold θ_k for this dichotomization.) This implies that the analyses of the dichotomized outcomes also provide insight into the fit of the model for the c categories. If the estimated regression parameters $\gamma_1, \dots, \gamma_r$ depend strongly on the dichotomization point, then it is likely that the multilevel multicategory logistic or probit model does not fit well.

Example 17.7 *The number of science subjects taken in high school.*

This example continues the analysis of Example 17.5, but now analyzes the number of science subjects instead of only whether this number was larger than 0.

The number of science subjects ranges from 0 to 3 (mathematics, chemistry, physics). There were 736 pupils who took no science subjects, 1,120 who took one, 873 who took two, and 703 who took all three.

The multilevel four-category logistic regression model was estimated by the MLwiN and MIXOR programs, with gender (0 for boys, 1 for girls), socio-economic status (an ordered scale with values 1–6, from which the mean, 3.68, was subtracted), and minority status (0 for children of parents born in industrialized countries, 1 for other countries) as explanatory variables. Both programs produced very similar results. The MIXOR results are presented so that the deviance is available.

Table 17.6 shows the result for the empty model (which only has the thresholds and the intercept variance as parameters) and for the model with these three explanatory variables. The deviance is minus twice the log-likelihood reported by MIXOR. Further, MIXOR reports standard deviations of random effects rather than variances. To obtain consistency with the other tables in this book, we report the intercept variance instead. The standard error of the variance estimate is obtained from the formula $S.E.(\hat{\sigma}^2) = 2\hat{\sigma} S.E.(\hat{\sigma})$, obtained by rearranging equation (6.2).

In the empty model the intercept variance is 0.243. In the model with explanatory variables this parameter has increased to 0.293. This increase is in accordance with what was said on p. 309 about the effect of including level-one variables in the model. The same effect is responsible for the fact that the threshold parameters have increased. All three fixed effects in Model 2 are significant: girls tend to take considerably fewer science subjects than boys, the number of science subjects is an increasing

Table 17.6: Multilevel four-category logistic regression model for the number of science subjects.

Threshold parameters	Model 1		Model 2	
	Threshold	S.E.	Threshold	S.E.
θ_1 Threshold 1–2	1.541	0.041	1.763	0.045
θ_2 Threshold 2–3	2.784	0.046	3.211	0.054
Fixed effect	Coefficient	S.E.	Coefficient	S.E.
γ_0 Intercept	1.370	0.057	2.591	0.079
γ_1 Gender girls			-1.680	0.066
γ_2 SES			0.117	0.037
γ_3 Minority status			-0.514	0.156
Level-two random effect	Var. comp.	S.E.	Var. comp.	S.E.
τ_0^2 Intercept variance	0.243	0.034	0.293	0.040
Deviance	9,308.8		8,658.2	

function of socio-economic status, and minority pupils tend to take fewer science subjects than non-minority pupils. The proportion of explained variance can be calculated from equation (17.26). The linear predictor is given here by

$$-1.680X_1 + 0.117X_2 - 0.514X_3,$$

and has variance 0.745. Hence the explained variance is

$$\frac{0.745}{0.745 + 0.293 + 3.29} = 0.17.$$

The thresholds are approximately equidistant (recall that the first threshold is $\theta_0 = 0$), with a difference between them in Model 2 of about 1.6. The gender effect has about the same size in absolute value, -1.68, so girls tend to take about one science subject fewer than boys. The effect of minority status, -0.514, amounts to an average difference of about 0.3 subjects. The random intercept standard deviation is $\sqrt{0.293} = 0.54$, so the difference between schools is quite large (as the span from schools with the few percent lowest U_{0j} to schools with the few percent highest U_{0j} is four standard deviations). The fact that the intercepts (fixed coefficients) are so different between Models 1 and 2 is because, of the explanatory variables, only SES is centered: for Model 1 the intercept corresponds to the average \bar{Y} value of all pupils, whereas for Model 2 it corresponds to the average \bar{Y} value of the nonminority boys.

17.5 Multilevel event history analysis

Event history analysis is the study of durations until some event occurs, such as recovery from a disease, entry into the labor market, or marriage; see, for example, Singer and

Willett (2003) and Mills (2011). When events are considered that can only happen once, and an approach is followed where time proceeds in discrete steps (days, years, etc.), a standard procedure is to transform the data into the so-called person-period format, which is a two-level format with periods nested within individuals, and with a binary outcome which is 0 if the event has not yet occurred, and 1 if it has occurred; and where for every person for whom the event has occurred during the observation period, only the first time step after the event is included in the data set. This transforms the outcome variable into a binary one, so that methods for binary response variables can be used. A single-level event history analysis can be modeled, for example, by logistic regression of such a data file. In spite of the two-level data structure this should be a single-level logistic regression without additional random effects, because the data for each person are restricted to being either a sequence of zeros if the event has never occurred, or a sequence of zeros followed by a single one if it has occurred. A multilevel nesting structure, with persons nested in groups (higher-level units), leads to a three-level data structure which can be modeled, for example, by multilevel logistic regression, with random effects for the groups, which then are the units at level three.

The logit link function leads here to the so-called proportional odds model for discrete time multilevel event history analysis, while the complementary log-log link function yields the proportional hazards model. These models, and the choice between them, are treated in the literature cited.

Multilevel models for event history analysis that fit within this framework are discussed by, for example, Steele (2008, Section 3), Rabe-Hesketh and Skrondal (2008, Chapter 8), Goldstein (2011), and Mills (2011, Chapter 8). The literature contains many examples; one of these, Windzio (2006), focuses on the incorporation of time-changing contextual variables. Other models for multilevel event history analysis were presented by Glidden and Vittinghoff (2004).

17.6 Multilevel Poisson regression

Another important type of discrete data is count data. For example, for a population of road crossings, one might count the number of accidents in one year; or, for a population of doctors, one could count how often in one year they are confronted with a certain medical condition. The set of possible outcomes of count data is the set of natural numbers: 0, 1, 2, 3, 4, The standard distribution for counts is the Poisson distribution – see textbooks on probability theory, or, for example, Long (1997, Section 8.1). A nice introduction to the analysis of count data on the basis of the Poisson distribution in cases where there are no explanatory variables can be found in Box et al. (1978, Section 5.6). An extensive treatment of Poisson and other models for count data is given by Cameron and Trivedi (1998).

The Poisson distribution is an approximation to the binomial distribution for the situation that the number of trials is large and the success probability is low. For the road accident example, one could consider a division of the year into 8,760 one-hour periods, and assume that the probability is negligible that more than one accident will happen in any single hour, that the occurrence of accidents is independent for different hours, and that the probability of an accident in any given hour, say, p , is very low. The total number

of accidents would then be binomially distributed with parameters 8,760 and p , and this distribution is extremely close to the Poisson distribution with mean $8,760 \times p$. Even when the ‘success’ probabilities are variable over trials (the accident probability is variable from one hour to the next), the number of ‘successes’ will still have approximately a Poisson distribution.

Just as we described in Section 17.1 for the binomial distribution, there is also for the Poisson distribution a natural relation between the mean and variance: they are equal. If Y has a Poisson distribution with mean λ , then

$$\mathcal{E}(Y) = \text{var}(Y) = \lambda. \quad (17.27)$$

If the counts tend to be large, their distribution can be approximated by a continuous distribution. If all counts are large enough (say, more than 8), then it is advisable to use the square root of the counts, \sqrt{Y} , as the dependent variable and apply the hierarchical linear model (with the usual assumption checks). The reason why this is a good approach resides in the fact that the square root transformation succeeds very well in transforming the Poisson distribution into an approximately homoscedastic normal distribution – the square root is the so-called variance-stabilizing transformation for the Poisson distribution (see Box et al., 1978, p. 144). An even better transformation to a homoscedastic distribution is the Freeman–Tukey transformation defined by $\sqrt{Y} + \sqrt{Y+1}$ (see Bishop et al., 1975, Section 14.6.2).

If all or some of the counts are small, a normal distribution will not be satisfactory and a hierarchical generalized linear model can be considered. This model is analogous to the multilevel logistic regression model. However, instead of the probability of success of Section 17.2 we now model the *expected value of the count*. Denote by Y_{ij} the count for level-one unit i in group j , and by L_{ij} the expected (i.e., population average) count, given that unit i is in group j and given the values of the explanatory variables (if any). Then L_{ij} is necessarily a nonnegative number, which could lead to difficulties if we considered linear models for this value (cf. Section 17.2.2). Just as the logit function was used as the link function for probabilities in Section 17.2.2, so the natural logarithm is mostly used as the link function for expected counts. For single-level data this leads to the Poisson regression model (Long 1997, Chapter 8) which is a linear model for the natural logarithm of the counts, $\ln(L_{ij})$. For multilevel data, hierarchical linear models are considered for the logarithm of L_{ij} . The random intercept Poisson regression model is thus formulated, analogously to (17.15), as a regression model plus a random intercept for the logarithm of the expected count:

$$\ln(L_{ij}) = \gamma_0 + \sum_{h=1}^r \gamma_h x_{hij} + U_{0j}. \quad (17.28)$$

The variance of the random intercept is again denoted by τ_0^2 . This model is treated in Diggle et al. (2002, Section 9.4), Hedeker and Gibbons (2006, Chapter 12), Goldstein (2011, Section 4.5), and Skrondal and Rabe-Hesketh (2004, Chapter 11).

The multilevel Poisson regression model can be estimated by various multilevel software packages (see Section 17.2.5 and Chapter 18), including glamm, HLM, MIX-PREG/SuperMix, MLwiN, various R packages including lme4 and glmmADMB, SAS, Stata, and Latent Gold. The estimation methods, as with those for multilevel logistic regression, include numerical integration (in the version of Gaussian quadrature, with or without

adaptive node placement; see Rabe-Hesketh et al., 2005), Laplace approximation (Raudenbush et al., 2000), and various Taylor series approximations (Goldstein, 2011). The numerical integration method and the Laplace approximation provide not only parameter estimates but also a deviance statistic which can be used for hypothesis testing.

To transform the linear model back to the expected counts, the inverse transformation of the natural logarithm must be used, which is the exponential function $\exp(x) = e^x$. This function has the property that it transforms sums into products:

$$e^{a+b} = e^a \times e^b.$$

Therefore the explanatory variables and the level-two random effects in the (additive) multi-level Poisson regression model have multiplicative effects on the expected counts. For example, if there is only $r = 1$ explanatory variable, equation (17.28) is equivalent to

$$\begin{aligned} L_{ij} &= \exp(\gamma_0 + \gamma_1 x_{1ij} + U_{0j}) \\ &= \exp(\gamma_0) \times \exp(\gamma_1 x_{1ij}) \times \exp(U_{0j}). \end{aligned} \quad (17.29)$$

Therefore, each additional unit of X_1 will have the effect of *multiplying* the expected count by e^{γ_1} . Similarly, in a group with a high intercept, for example, two standard deviations so that $U_{0j} = 2\tau_0$, the expected count will be $e^{2\tau_0}$ times as high as in a group with an average value, $U_{0j} = 0$, of the intercept.

In models for counts it is quite usual that there is a variable D that is known to be proportional to the expected counts. For example, if the count Y_{ij} is the number of events in some time interval of nonconstant length d_{ij} , it often is natural to assume that the expected count is proportional to this length of the time period. If in the example of counts of medical problems there are several doctors each with his or her own population of patients, then D could be the size of the patient population of the doctor. In view of equation (17.29), in order to let the expected count be proportional to D , there should be a term $\ln(d_{ij})$ in the linear model for $\ln(L_{ij})$, with a regression coefficient fixed to 1. Such a term is called an *offset* in the linear model (see McCullagh and Nelder, 1989). Goldstein (2011, Section 4.5) suggests that such offset variables be centered to improve the numerical properties of the estimation algorithm.

Example 17.8 Memberships in voluntary organizations.

Memberships in voluntary organizations may be regarded as a measurement of social activity and a dimension of social capital. In this example a subset is used of the data also analyzed in an international comparative study by Peter and Drobnić (2010). Here only the data for the Netherlands are studied. They were collected in the Dutch part of the European Social Survey (ESS), in the ‘Citizenship, Involvement and Democracy’ module, during 2002–2003.

The dependent variable is the number of types of association of which the respondent is a member. Twelve types of organizations are mentioned, ranging from sports clubs through political parties to consumer organizations. Thus, the count can range from 0 to 12. This data set has a regional identifier which ranges over 40 regions for the Netherlands. Respondents aged between 25 and 65 years were selected, and a very small number with missing values were discarded. This left a total of 1,738 respondents. The nesting structure is respondents nested in regions.

The number of memberships ranges from 0 to 10, with a mean of 2.34 and a standard deviation of 1.78. The variance is 3.16. For the Poisson distribution according to (17.27) in the population the mean and variance are equal. The fact that here dispersion is greater than would be expected for the Poisson distribution points to heterogeneity, which may have to do with individual differences and/or differences between regions. The Poisson distribution can assume all nonnegative values while the

outcome variable here is limited by the questionnaire to a maximum of 12, but in view of the low mean and standard deviation this in itself is not a problem for applying the Poisson distribution.

Deviating from our practice in earlier chapters, here we report for the random part the standard deviations and correlation parameters, rather than the variances and covariance. There are two reasons for this. One is the fact that the numbers reported are smaller. The other is the use of the R packages lme4 and glmmADMB for the computations, which use these parameters for reporting.

The empty model is reported as Model 1 in Table 17.7. Explorations showed that religion (coded as Protestant versus other), gender, and age had effects, and the effect of age was well described by a quadratic function. Furthermore, there appeared to be a random slope for gender and an interaction between age and gender. This is reported as Model 2. Age is centered at 40 years. To obtain parameters that are not too small, age is measured in decades: for example, 30 years is coded as -1 , 40 years as 0 , 50 years as $+1$.

Table 17.7: Two Poisson models for number of memberships.

Fixed effect	Model 1		Model 2	
	Coefficient	S.E.	Coefficient	S.E.
γ_0 Intercept	0.846	0.026	0.860	0.032
γ_1 Female			-0.118	0.044
γ_2 $(\text{Age} - 40)/10$			0.198	0.028
γ_3 $(\text{Age} - 40)^2/100$			-0.061	0.014
γ_4 Protestant			0.296	0.041
γ_5 Female $\times (\text{Age} - 40)/10$			-0.097	0.030
Level-two random part	S.D.		S.D. / Corr.	
τ_0 Intercept standard deviation	0.112		0.070	
τ_1 Slope S.D. Female			0.146	
$\rho_{01}(\tau)$ Int.–slope correlation			0.168	
Deviance	6,658.3		6,522.6	

A random intercept model with the same fixed effects as Model 2 has deviance 6531.9. The deviance difference is $6531.9 - 6522.6 = 9.3$. Against a chi-squared distribution with 2 degrees of freedom this has $p < 0.01$, even without the more powerful test (cf. Section 6.2.1). This shows there is clear evidence for a random slope for gender: the effect of gender on number of membership differs between regions.

We see that Protestants have more memberships on average. To assess the contribution of gender, we must take into account that there is an interaction between gender and age, as well as a random slope for gender. Since age is centered at 40 years, the main effect of gender as well as the intercept variance refer to age 40. It can be concluded that on average, females have fewer memberships at 40 years. The total contribution of being female is

$$-0.118 - 0.097(\text{Age} - 40)/10 + U_{1j},$$

which is 0 for age 28 if $U_{1j} = 0$. It can be concluded that in an average region, females have on average the same number of memberships as males at age 28, and the average increase by age is

smaller for females than for males. But the male–female difference depends considerably on region, as the between-region standard deviation is $S.D.(U_{1j}) = \sqrt{0.0211} = 0.15$, larger than the main effect of gender. Therefore there are many regions where at age 40 females have more memberships than males.

In linear regression we are accustomed to not being very concerned about deviations from normality for the distribution of residuals, as long as there are no serious outliers. For Poisson regression this robustness to deviations from the distributional assumption does not hold. The reason for the difference is that the normal distribution has two separate parameters for the mean and variance, whereas the Poisson distribution has just one parameter, and in formula (17.27) we saw that the mean and variance are equal. It is not unusual, however, for counts to have a distribution where the variance differs from the mean; in most cases the variance will then be larger than the mean, which is called *overdispersion* with respect to the Poisson distribution. To obtain robustness for such deviations from the Poisson distribution, two approaches can be used.

One is to add overdispersion to the model without precisely specifying the resulting probability distribution (McCullagh and Nelder, 1989). Denote the overdispersion parameter by φ ; this must be positive. Given that the mean of the distribution is parametrized by λ just like the Poisson distribution, the mean and variance of the distribution are then given by

$$\mathcal{E}(Y) = \lambda \quad \text{var}(Y) = \varphi \lambda. \quad (17.30)$$

Thus, $\varphi = 1$ represents the Poisson distribution, $\varphi > 1$ represents overdispersion, and $\varphi < 1$ underdispersion, which is possible but less common.

The other approach is to replace the Poisson distribution by the so-called negative binomial distribution (see, for example, Long, 1997). The negative binomial distribution has two parameters, $\lambda \geq 0$ and $\alpha > 0$. The only property we need to know of this distribution is that if the random variable Y has a negative binomial distribution with parameters λ and α , then its mean and variance are given by

$$\mathcal{E}(Y) = \lambda, \quad \text{var}(Y) = \lambda \left(1 + \frac{\lambda}{\alpha}\right). \quad (17.31)$$

Thus, the parameter α determines the amount of overdispersion; the smaller it is, the greater the overdispersion; and as α grows very large the Poisson distribution is obtained as a boundary case.

The two mean–variance relationships (17.30) and (17.31) are different, because the latter can represent only overdispersion and the variance-to-mean ratio also increases as a function of the mean λ , whereas the former can represent underdispersion as well, and has a constant variance-to-mean ratio. Which of the two relationships is more suitable is an empirical matter, but the practical interpretation of the results for an overdispersed Poisson model and for a negative binomial model will usually be very similar.

Thus, for a dependent variable which is a count, or more generally a nonnegative integer, one can choose between Poisson regression and negative binomial regression. The latter is more robust and has Poisson regression as a boundary case, and is therefore preferable unless the researcher is quite confident that the dependent variable, given the predictor variables and given the random effects, indeed has a Poisson distribution. Poisson regression with an overdispersion parameter often leads to quite similar conclusions to negative

binomial regression; it has the disadvantage that it is not based on an explicit probability distribution, but the advantage that it also allows underdispersion.

Further treatments of Poisson, overdispersed Poisson, and negative binomial regression are given by Gelman and Hill (2007, Chapter 15) and Berk and MacDonald (2008). Berk and MacDonald warn that it is important especially to work toward a good specification of the predictors and covariance structure of the model, rather than resorting too easily to overdispersed modeling as quick fix. A mathematical treatment is given by Demidenko (2004), who also provides a test for overdispersion (Section 7.5.10).

Example 17.9 Negative binomial regression for memberships in voluntary organizations.

We continue the example on memberships in voluntary organizations, now applying negative binomial regression. The same models are estimated as in Table 17.7, but now for negative binomial regression. The results are presented in Table 17.8.

Table 17.8: Two negative binomial models for number of memberships.

Fixed effect	Model 3		Model 4	
	Coefficient	S.E.	Coefficient	S.E.
γ_0 Intercept	0.848	0.027	0.861	0.034
γ_1 Female			-0.116	0.046
$\gamma_2 (Age - 40)/10$			0.197	0.031
$\gamma_3 (Age - 40)^2/100$			-0.061	0.016
γ_4 Protestant			0.300	0.047
γ_5 Female \times $(Age - 40)/10$			-0.095	0.033
Level-two random part	Parameter		Parameter	
τ_0 Intercept standard deviation	0.111		0.051	
τ_1 Slope S.D. Female			0.127	
$\rho_{01}(\tau)$ Int.–slope correlation			0.497	
α Negative binomial parameter	7.40		10.32	
Deviance	6,588.6		6,483.3	

Comparing the two empty models, Model 1 (Table 17.7) and Model 3, we see that Model 3 fits better (deviance difference $6,658.3 - 6,588.6 = 69.7$, $df = 1$, $p < 0.001$), the negative binomial parameter is rather large ($\hat{\alpha} = 7.4$), indicating a moderate degree of overdispersion, the estimated intercept parameter is virtually the same, and the intercept standard deviation is slightly less in Model 3 than in Model 1. The overdispersion occurs at level one, but in Model 1 it cannot be represented anywhere at level one and therefore adds to the dispersion of the level-two random effects.

Comparing the two substantive models, Model 2 and Model 4, we see again that Model 4 fits better (deviance difference $6,522.6 - 6,483.3 = 379.3$, $df = 1$, $p < 0.001$), all regression coefficients are practically the same, the standard deviations of the intercept and slope are smaller in Model 4 than in Model 2, and the slope–intercept correlation is larger. The standard errors of the fixed effects are all slightly larger in Model 4 than in Model 2. The negative binomial parameter has increased compared to Model 3 ($\hat{\alpha} = 10.3$), which shows that apparently the explanatory variables capture some, but not all, of what appears to be overdispersion in the empty model.

The larger standard errors in Model 4 are definitely more trustworthy than those in Model 2, given our observations above about the greater robustness of overdispersed models. Similarly, the different parameters for the level-two random effects in Model 4 are more trustworthy than those in Model 2. In this case the differences between the overdispersed and the Poisson models are moderate, but in other data sets they may be more important.

17.7 Glommary

Dichotomous variables. Also called binary variables, these are variables with two possible values, such as ‘yes’ and ‘no’; often the values are formally called ‘success’ and ‘failure’. For such dependent variables, a hierarchical linear model with homoscedastic normally distributed residuals is not appropriate, in the first place because the residual variance of a binary dependent variable will depend on the predicted value. The same holds for dependent variables with a small number of ordered numerical categories, because the residual variance must become small when the predicted value approaches either of the extremes of the range of values; and for nonnegative integer (count) data, because the residual variance must become small when the predicted value approaches 0.

Hierarchical generalized linear model. This is similar to the hierarchical linear model, but for dependent variables that are not normally distributed given the explanatory variables. For dichotomous variables, the best-known hierarchical generalized linear model is the multilevel logistic regression model. In this model, the success probability is the logistic transform of the linear predictor – the fixed part of the model – plus the contribution made by the random effects. The latter will usually comprise a random intercept to which random slopes may be added; as in the regular hierarchical linear model, these random coefficients vary over the level-two units.

Logistic function. This is given by $\text{logistic}(x) = \exp(x) / (1 + \exp(x))$, and maps the set of real numbers on to the interval $(0, 1)$.

Logit function. This is given by $\text{logit}(x) = \ln(p/(1 - p))$, and maps the interval $(0, 1)$ on to the set of real numbers. It is also called the log-odds. It is used as the link function in logistic regression, linking the linear predictor to the expected value of the dependent variable. The logit function is the inverse of the logistic function, that is, $\text{logistic}(\text{logit}(p)) = p$ and $\text{logit}(\text{logistic}(x)) = x$.

Empty model for dichotomous data. This is a logistic model for dichotomous data where the linear predictor is constant (consisting of only the fixed intercept) and the logit of the probability of success is a random intercept, that is, the constant term plus a random residual depending on the level-two unit.

Random intercept model for dichotomous data. This is a logistic model for dichotomous data where the linear predictor consists of a linear combination of explanatory variables, and in addition there is a random intercept.

Random slope model for dichotomous data. In addition to the linear predictor and the random intercept, this comprises a random slope for one or more level-one variables.

Multilevel logistic regression model. This is a logistic model with random coefficients, also called the multilevel logit model.

Multilevel probit model. This is quite analogous to the multilevel logistic regression model, but now the link function, transforming the linear predictor plus random coefficients to the probability of success, is the cumulative distribution function of the standard normal distribution.

Threshold representation. This is a nice way of interpreting the multilevel logit and probit models. It postulates a continuous variable, unobserved, of which the distribution is the linear predictor plus a ‘residual’ having, respectively, a logistic or a standard normal distribution, and of which only the sign (positive = success, negative = failure) is observed.

Intraclass correlation for binary variables. This can be defined in two ways. First, it can be defined just as for other variables, as in Chapter 3. Second, it can be defined in terms of the underlying continuous variable in the threshold representation. The second definition has the advantage that it corresponds to an elegant definition of the proportion of explained variance (see below).

Testing that the intraclass correlation for binary variables is 0. This can be done by applying the well-known chi-squared test. Another test was also mentioned, which can always be applied when there are many groups (even if they are small, which may lead to difficulties for the chi-squared test) as long as there is not a very small number of groups making up almost all of the data.

Proportion of explained variance for dichotomous dependent variables. This can be defined as the proportion of explained variance for the underlying continuous variable in the threshold representation.

Adding effects of level-one variables. Doing this to a multilevel logistic regression model can increase the random intercept variance and the effects of uncorrelated level-one variables; this was explained by considering the threshold representation.

Multilevel ordered logistic regression model. This is a hierarchical generalized linear model for dependent variables having three or more ordered categories. It can be represented as a threshold model just like the multilevel logistic regression model for dichotomous outcome variables, with $c - 1$ thresholds between adjacent categories, where c is the number of categories.

Event history analysis. The study of durations until some event occurs. One approach is to transform data to the person-period format, which is a two-level format with time periods nested within individuals, and with a binary outcome which is 0 if the event has not yet occurred, and 1 if it has occurred; including only the first (if any) 1 response per individual. A multilevel nesting structure, with persons nested in groups (higher-level units), is then represented by a three-level data structure (periods within persons within groups) which can be modeled by multilevel logistic regression, with random effects at level three.

Multilevel Poisson regression. A hierarchical generalized linear model for count data, in which the dependent variable has possible values 0, 1, 2, Conditional on the explanatory variables and the random effects, under this model the dependent variable has a Poisson distribution. This is the most often used probability distribution for counts. A particular property of the Poisson distribution is that the variance is equal to the expected value. If the expected values are large (say, all larger than 8) then an alternative approach is to apply a square root transformation to the dependent variable and use the regular hierarchical linear model.

Overdispersion. For count data, overdispersion means that, given the explanatory variables and random effects, the residual variance is larger than the expected value. If this is the case, the use of the Poisson distribution may lead to erroneous inferences – in particular, underestimated standard errors. Then it is advisable to use negative binomial regression. The negative binomial distribution is another distribution for count data, but it has an extra parameter to accommodate variances larger than the mean. Another option is to use a multilevel Poisson regression model to which an overdispersion parameter has been added that does not correspond to a particular probability distribution.

Estimation methods. For hierarchical generalized linear models, estimation is more complicated than for the hierarchical linear model with normal distributions. Various algorithms (i.e., procedures for calculating estimates) are used. One type of estimation procedure is maximum likelihood; for these models approximations to the maximum likelihood estimator have been developed using adaptive or nonadaptive quadrature (i.e., numerical integration) and the Laplace approximation. Bayesian methods have been developed based on simulation (Markov chain Monte Carlo) and on the Laplace approximation. Marginal quasi-likelihood and penalized quasi-likelihood are estimation procedures which were developed earlier and which are less effective compared to maximum likelihood estimation by numerical integration or to Bayesian methods. For the estimation of fixed effects when group sizes are not too small, however, these methods still can perform quite well.

18

Software

Almost all procedures treated in this book can be carried out by standard software for multilevel statistical models. This of course is intentional, since this book covers those parts of the theory of the multilevel model that can be readily applied in everyday research practice. However, things change rapidly. Some of the software discussed in the previous edition of this book is no longer available. On the other hand, new software packages are shooting up like mushrooms. The reader is therefore advised to keep track of the changes that can be found at the various websites we mention in this chapter – although we have to add that some of these websites and even internet addresses tend to change. But then again, search engines will readily guide the curious multilevel researcher to the most recent sites.

Currently most details on the specialized multilevel software packages can be found via the links provided on the homepage of the Centre for Multilevel Modelling at the University of Bristol; see <http://www.bristol.ac.uk/cmm/>. At this site one can also find reviews of the multilevel software packages. Chapter 18 of Goldstein (2011) also provides a list of computer programs for multilevel analysis.

OVERVIEW OF THE CHAPTER

We will provide a brief review of the multilevel software packages, organized into three sections. Section 18.1 is devoted to special purpose programs, such as HLM and MLwiN, specifically designed for multilevel modeling. Section 18.2 treats modules in general-purpose software packages, such as R, SAS, Stata, and SPSS, that allow for multilevel modeling. Finally, Section 18.3 mentions some specialized software programs, built for specific research purposes.

18.1 Special software for multilevel modeling

Multilevel software packages are aimed at researchers who specifically seek to apply multilevel modeling techniques. Other statistical techniques are not available in these specific multilevel programs (with some exceptions). Each of the programs described in this section was designed by pioneers in the field of multilevel modeling. We only review HLM,

MLwiN and the MIXOR suite, although there are other packages as well. The interested reader may take a look at [http://www.bristol.ac.uk/cmm/learning/mmsoftware/](http://www.bristol.ac.uk/cmm/learning/mmssoftware/) to find descriptions and reviews of other multilevel software programs such as aML, EGRET, and GENSTAT, stemming from the fields of econometrics, epidemiology, and agriculture, respectively.

18.1.1 HLM

HLM was originally written by Bryk et al. (1996), and the theoretical background behind most applications can be found in Raudenbush and Bryk (2002). The main features of HLM are its interactive operation (although one can also run the program in batch mode) and the fact that it is rather easy to learn. Therefore it is well suited for undergraduate courses and for postgraduate courses for beginners. The many options available also make it a good tool for professional researchers. Information is obtainable from the website, <http://www.ssicentral.com/hlm/>, which also features a free student version. West et al. (2007) give an introduction to hierarchical linear modeling with much attention to implementation in HLM.

Input consists of separate files for each level in the design, linked by common identifiers. In a simple two-level case, for example, with data about students in schools, one file contains all the school data with a school identification code, while another file contains all the student data with the school identification code for each student. The input can come from system files of SPSS, SAS, SYSTAT or Stata, or may be given in the form of ASCII text files.

Once data have been read and stored into a sufficient statistics file (a kind of system file), there are three ways to work with the program. One way is to run the program interactively (answering questions posed by the program). Another is to run the program in batch mode. Batch and interactive modes can also be combined. Finally, one can make full use of the graphical interface. In each case the two-step logic of Section 6.4.1 is followed. HLM does not allow for data manipulation, but both the input and output can come from, and can be fed into, SPSS, SAS, SYSTAT, or Stata. HLM does not go beyond four levels. It can be used for practically all the analyses presented in this book, with the exception of multiple membership models. Almost all examples in this book can be reproduced using HLM version 7. Some interesting features of the program are the ability to test model assumptions directly, for example, by test (10.5) for level-one heteroscedasticity, and the help provided to construct contrast tests. Furthermore, the program routinely asks for centering of predictor variables, but the flipside of the coin is – if one opts for group mean centering – that group means themselves must have been calculated outside HLM, if one wishes to use these as level-two predictor variables.

A special feature of the program is that it allows for statistical meta-analysis (see Section 3.7) of research studies that are summarized by only an effect size estimate and its associated standard error (called in Raudenbush and Bryk, 2002, the ‘V-known problem’). Other special features are the analysis of data where explanatory variables are measured with error (explained in Raudenbush and Sampson, 1999a), and the analysis of multiply imputed data as discussed in Chapter 9. Note, however, that the imputations have to be done with other specialized software packages. Next to that HLM also offers facilities for the analysis of a special case of multiply imputed data, namely the multilevel analysis of plausible values. These are imputed data on the dependent variable for incomplete designs,

such as designs with booklet rotations as used in the PISA studies. As a special feature it allows for the modeling of dependent random effects, a topic not treated in this book.

18.1.2 MLwiN

MLwiN is the most extensive multilevel package, written by researchers currently working at the Centre for Multilevel Modelling at the University of Bristol (Rasbash and Woodhouse, 1995; Goldstein et al., 1998; Rasbash et al., 2009). Current information, including a wealth of documentation, can be obtained from the website <http://www.bristol.ac.uk/cmm/>. Almost all the examples in this book can be reproduced using MLwiN, which allows for standard modeling of up to five levels. For heteroscedastic models (Chapter 8), the term used in the MLwiN documentation is ‘complex variation’. For example, level-one heteroscedasticity is complex level-one variation. Next to the standard IGLS (ML) and RIGLS (REML) estimation methods, MLwiN also provides bootstrap methods (based on random drawing from the data set or on random draws from an estimated population distribution), and extensive implementation of Markov chain Monte Carlo methods (Browne, 2009) for Bayesian estimation (see Section 12.1). With MLwiN one may use the accompanying package REALCOM-Impute to impute missing data in a multilevel model, and analyze these subsequently in MLwiN. The program MLPowSim (Section 18.3.3) may be used to create MLwiN macros for simulation-based power analysis.

A nice feature of MLwiN is that the program was built on NANOSTAT, a statistical environment which allows for data manipulation, graphing, simple statistical computations, file manipulation (e.g., sorting), etc. Data manipulation procedures include several handy procedures relating to the multilevel data structure. Input for MLwiN may be either an ASCII text file or a system file from Minitab, SPSS, or Stata that contains all the data, including the level identifiers. MLwiN can even be called directly from Stata using the runmlwin command (Leckie and Charlton, 2011). Since the data are available in one file, this implies that all the level-two and higher-level data are included in disaggregated form at level one. The data are read into a worksheet, a kind of system file. This worksheet, which can also include model specifications, variable labels, and results, can be saved and used in later sessions. The data can also be exported to Minitab, SPSS, or Stata.

The most obvious way to work with the program is interactively using the graphical interface. One can, however, also use the ‘command menu’ or give a series of commands in a previously constructed macro.

MLwiN is the most flexible multilevel software package, but it may take some time to get acquainted with its features. It is an excellent tool for professional researchers and statisticians. The macro facilities in particular provide experienced researchers ample opportunities for applications such as meta-analysis, multilevel factor analysis, multilevel item response theory modeling, to name just some examples.

18.1.3 The MIXOR suite and SuperMix

Hedeker and Gibbons (1996a, b) have constructed several modules for multilevel modeling, focusing on special models that go beyond the basic hierarchical linear model. DOS and Windows versions of these programs are freely available with manuals and examples at <http://tigger.uic.edu/~hedeker/mixwin.html>. These programs have no facilities for data

manipulation and use ASCII text files for data input. The algorithms are based on numerical integration (see Section 17.2.5).

MIXREG (Hedeker and Gibbons, 1996b) is a computer program for mixed effects regression analysis with autocorrelated errors. This is suitable especially for longitudinal models (see Chapter 15). Various correlation patterns for the level-one residuals R_{ti} in (15.32) are allowed, such as autocorrelation or moving average dependence.

MIXOR (Hedeker and Gibbons, 1996a) provides estimates for multilevel models for dichotomous and ordinal discrete outcome variables (cf. Chapter 17). It allows probit, logistic, and complementary log-log link functions. These models can include multiple random slopes. It also permits right-censoring of the ordinal outcome (useful for analysis of multilevel grouped-time survival data), nonproportional odds (or hazards) for selected covariates, and the possibility for the random effect variance terms to vary by groups of level-one or level-two units. This allows estimation of many types of item response theory models (where the variance parameters vary by the level-one items) as well as models where the random effects vary by groups of subjects (e.g., males versus females).

MIXNO implements a multilevel multinomial logistic regression model. As such, it can be used to analyze two-level categorical outcomes without an ordering. As in MIXOR, the random effect variance terms can also vary by groups of level-one or level-two units. This program has an extensive manual with various examples.

MIXPREG is a program for estimating the parameters of the multilevel Poisson regression model (see Section 17.6). This program also has an extensive manual with examples.

Finally, the program SuperMix combines all these packages with the multilevel model for continuous data as an add-on, and moreover provides ample opportunities for data manipulation, importing of files from statistical packages or Excel, and allows the user to specify three-level models using a graphical interface. This package is available at <http://www.ssicentral.com/supermix/>, where one also can download a free student version with limited capacity.

In the interpretation of the output of these programs, it should be noted that the parametrization used for the random part is not the covariance matrix but its Cholesky decomposition. This is a lower triangular matrix C with the property that $CC' = \Sigma$, where Σ is the covariance matrix. The output does also give the estimated variance and covariance parameters, but (in the present versions) not their standard errors. If you want to know these standard errors, some additional calculations are necessary.

If the random part only contains the random intercept, the parameter in the Cholesky decomposition is the standard deviation of the intercept. The standard error of the intercept variance can be calculated with formula (6.2). For other parameters, the relation between the standard errors is more complicated. We treat only the case of one random slope. The level-two covariance matrix is then

$$T = \begin{pmatrix} \tau_0^2 & \tau_{01} \\ \tau_{01} & \tau_1^2 \end{pmatrix}$$

with Cholesky decomposition denoted by

$$C = \begin{pmatrix} c_{00} & 0 \\ c_{01} & c_{11} \end{pmatrix}.$$

The correspondence between these matrices is

$$\begin{aligned}\tau_0^2 &= c_{00}^2, \\ \tau_{01} &= c_{00} c_{01}, \\ \tau_1^2 &= c_{01}^2 + c_{11}^2.\end{aligned}$$

Denote the standard errors of the estimated elements of C by s_{00} , s_{01} , and s_{11} , respectively, and the correlations between these estimates by $r_{00,01}$, $r_{00,11}$, and $r_{01,11}$. These standard errors and correlations are given in the output of the programs. Approximate standard errors for the elements of the covariance matrix of the level-two random part are then given by

$$\begin{aligned}\text{S.E.}(\hat{\tau}_0^2) &\approx 2c_{00}s_{00}, \\ \text{S.E.}(\hat{\tau}_{01}) &\approx \sqrt{c_{00}^2 s_{01}^2 + c_{01}^2 s_{00}^2 + 2c_{00}c_{01}s_{00}s_{01}r_{00,01}}, \\ \text{S.E.}(\hat{\tau}_1^2) &\approx 2\sqrt{c_{01}^2 s_{01}^2 + c_{11}^2 s_{11}^2 + 2c_{01}c_{11}s_{01}s_{11}r_{01,11}}.\end{aligned}$$

For the second and further random slopes, the formulas for the standard errors are even more complicated. These formulas can be derived with the multivariate delta method, explained, for example, in Bishop et al. (1975, Section 14.6.3).

18.2 Modules in general-purpose software packages

Several of the main general-purpose statistical packages have incorporated modules for multilevel modeling. These are usually presented as modules for mixed models, random effects, random coefficients, or variance components. As it is, most researchers may be used to one of these packages and may feel reluctant to learn to handle the specialized software discussed above. For these people especially the threshold to multilevel modeling may be lowered when they can stick to their own software. We provide a brief introduction. For further details the interested reader may turn to Twisk (2006), who discusses the possibilities of SPSS, Stata, SAS, and R, and also gives some examples of setups and commands to run multilevel models. And of course the website of the Centre for Multilevel Modelling contains reviews of the options the general-purpose packages provide for multilevel analysis as well; see [http://www.bristol.ac.uk/cmm/learning/mmsoftware/](http://www.bristol.ac.uk/cmm/learning/mmssoftware/). West et al. (2007) give a treatment of the hierarchical linear model (but not of generalized linear versions) with much attention to the use of HLM, SAS, SPSS, and Stata.

18.2.1 SAS procedures VARCOMP, MIXED, GLIMMIX, and NLMIXED

The SAS procedure MIXED has been up and running since 1996 (Littell et al., 2006). For experienced SAS users a quick introduction to multilevel modules in general-purpose software packages modeling using SAS can be found in Singer (1998a, 1998b). Verbeke and Molenberghs (1997) present a SAS-oriented practical introduction to the hierarchical linear model. The book on generalized hierarchical linear models for longitudinal data by Molenberghs and Verbeke (2006) also offers much explanation of the implementation

in SAS. An introduction to using SAS proc MIXED for multilevel analysis, specifically geared at longitudinal data, is given by Singer (2002).

Unlike some other general-purpose packages, SAS allows quite general variance components models to be fitted with VARCOMP, and in MIXED one can fit very complex multilevel models and calculate all corresponding statistics (such as deviance tests). MIXED is a procedure oriented toward general mixed linear models, and can be used to analyze practically all the hierarchical linear model examples for continuous outcome variables presented in this book. The general mixed model orientation has the advantage that crossed random coefficients can be easily included, but the disadvantage is that this procedure does not provide the specific efficiency for the nested random coefficients of the hierarchical linear model that is provided by dedicated multilevel programs. Also available are the procedures GLIMMIX and NLMIXED, which can be used to fit the models for discrete outcome variables described in Chapter 17.

18.2.2 R

R is an open source language and environment for statistical computing and graphics similar to the S language. It is highly flexible and freely downloadable at <http://cran.r-project.org/>. Due to the open source character of R one can use procedures that have been developed by others (Ihaka and Gentleman, 1996). Procedures in R are organized into so-called packages. R is operated by a command language, and the commands are collected in scripts. In the initial phase it may require some effort to learn the command language, but once mastered this has the advantage of flexibility and reproducibility of results.

There are several packages in R implementing multilevel models. The main ones are nlme (Pinheiro and Bates, 2000) and lme4 (Bates, 2010; see also Doran et al., 2007). The nlme package has extensive possibilities for linear and nonlinear models with normally distributed residuals. The lme4 package is currently still under vigorous development and, in spite of its title ('Linear mixed-effects models using S4 classes') also estimates hierarchical generalized linear models, using Laplace and other methods. With both of these packages one can perform most of the analyses treated in this book. Some texts introducing multilevel analysis using R are, next to the two books just mentioned, Maindonald and Braun (2007, Chapter 10), Bliese (2009), Wright and London (2009), Berridge and Crouchley (2011), and, specifically for multilevel models for discrete data, Thompson (2009, Chapter 12).

There are several packages that can be used for more limited, but very useful, purposes. The R packages mlmmm and mice allow multiple imputation of missing data (see Chapter 9) under a two-level model. The pamm package contains procedures for simulation-based power analysis. The program MLPowSim (Section 18.3.3) creates R scripts for simulation-based power analysis. Various new methods for hierarchical generalized linear models (Section 17.1) have been made available in R; some of these packages are HGLMM, glmmml, glmmADMB, and INLA. Package multilevel contains some procedures that are especially useful for those working with multi-item scales. Using packages WinBUGS or glmmBUGS gives access to the WinBUGS program (Section 18.3.7).

18.2.3 Stata

Stata (StataCorp, 2009) contains some modules that permit the estimation of certain multilevel models (see Rabe-Hesketh and Skrondal, 2008). Module loneway ('long oneway')

gives estimates for the empty model. The xt series of modules are designed for the analysis of longitudinal data (cf. Chapter 15), but can be used to analyze any two-level random intercept model. Command xtreg estimates the random intercept model, while xtpred calculates posterior means. Commands xtpois and xtprobit, respectively, provide estimates of the multilevel Poisson regression and multilevel probit regression models (Chapter 17). These estimates are based on the so-called generalized estimating equations method. A special feature of Stata is the so-called sandwich variance estimator, also called the robust or Huber estimator (Section 12.2). This estimator can be applied in many Stata modules that are not specifically intended for multilevel analysis. For statistics calculated in a single-level framework (e.g., estimated OLS regression coefficients), the sandwich estimator, when using the keyword ‘cluster’, computes standard errors that are asymptotically correct under two-stage sampling. In terms of our Chapter 2, this solves many instances of ‘dependence as a nuisance’, although it does not help to get a grip on ‘interesting dependence’.

Within Stata one can use the procedure GLLAMM – an acronym for General Linear Latent And Mixed Models – which was developed by Rabe-Hesketh et al. (2004, 2005) and which analyzes very general models indeed, including those of Chapter 17 but also much more general models with latent variables, such as multilevel structural equation modeling and multilevel item response theory models. The algorithms use adaptive quadrature (Section 17.2.5). The package is available at <http://www.gllamm.org>, where one can also find the necessary documentation.

As mentioned above, from Stata one may call MLwiN through the command runmlwin.

18.2.4 SPSS, commands VARCOMP and MIXED

The simplest introduction to multilevel modeling for SPSS users is the module (in the language of SPSS, ‘command’) VARCOMP. One can get no further, however, than the random intercept and random slope model, as described in Chapters 4 and 5 of this book, with the possibility also of including crossed random effects (Chapter 13). The SPSS module MIXED goes further and provides ample opportunities for three-level models with random slopes and complex covariance structures. A detailed introduction into multilevel modeling with SPSS is provided by Heck et al. (2010).

18.3 Other multilevel software

There are other programs available for special purposes, and which can be useful to supplement the software mentioned above.

18.3.1 PinT

PinT is a specialized program for calculations of *Power in two-level designs*, implementing the methods of Snijders and Bosker (1993). This program can be used for *a priori* estimation of standard errors of fixed coefficients. This is useful in the design phase of a multilevel study, as discussed in Chapter 11. Being shareware, it can be downloaded with the manual (Bosker et al., 2003) from <http://www.stats.ox.ac.uk/~snijders/multilevel.htm#progPINT>.

18.3.2 Optimal Design

Optimal Design is a flexible program for designing multilevel studies and was developed by Spybrook et al. (2009). The designs considered include cluster randomized trials, multisite trials, and repeated measures. The program is very flexible and even allows for power calculations for single-level randomized trials, cluster randomized trials with the outcome of interest at the cluster level, three-level experiments, and meta-analysis. The graphical interface is very user-friendly, and within a short while one can manipulate significance levels, number of clusters, number of observations per cluster, intraclass correlations, associations between covariates and outcomes, and infer effects on the resulting power. The software and the manual can be downloaded from http://www.wtgrantfoundation.org/resources/overview/research_tools.

18.3.3 MLPowSim

MLPowSim (Browne et al., 2009) is a program for power analysis and determination of sample sizes for quite general random effect models. Standard errors and power are computed based on Monte Carlo simulations for which either R or MLwiN can be used. MLPowSim creates R command scripts and MLwiN macro files which, when executed in those respective packages, employ their simulation facilities and random effect estimation engines to perform these computations.

18.3.4 Mplus

Mplus is a program with very general facilities for covariance structure analysis (Muthén and Muthén, 2010). Information about this program is available at <http://www.StatModel.com>. This program allows the analysis of univariate and multivariate two-level data not only with the hierarchical linear model but also with path analysis, factor analysis, and other structural equation models. Introductions to this type of model are given by Muthén (1994) and Kaplan and Elliott (1997).

18.3.5 Latent Gold

Latent Gold (Vermunt and Magidson, 2005a, 2005b) is a program primarily designed for analyzing latent class models, including the multilevel latent class models discussed in Section 12.3. It can also estimate the more regular multilevel models treated in other chapters. This program is available at http://www.statisticalinnovations.com/products/latentgold_v4.html.

18.3.6 REALCOM

At the time of writing the people at the Centre for Multilevel Modelling are developing new software for, as they call it, ‘realistic multilevel modeling’. Such modeling includes measurement errors in variables, simultaneous outcomes at various levels in the hierarchy, structural equation modeling, modeling with imputed data for missing values, and modeling with misclassifications. The interested reader is referred to <http://www.bristol.ac.uk/cmm/software/realcom> for more details and recent developments.

18.3.7 WinBUGS

A special program which uses the Gibbs sampler is WinBUGS (Lunn et al., 2000), building on the previous BUGS program (Gilks et al., 1996). Gibbs sampling is a simulation-based procedure for calculating Bayesian estimates (Section 12.1). This program can be used to estimate a large variety of models, including hierarchical linear models, possibly in combination with models for structural equations and measurement error. It is extremely flexible, and used, for example, as a research tool by statisticians; but it can also be used for regular data analysis. Gelman and Hill (2007) and Congdon (2010) give extensive attention to the use of WinBUGS for multilevel analysis. The WinBUGS example manuals also contain many examples of hierarchical generalized linear models.

The program is available with manuals from <http://www.mrc-bsu.cam.ac.uk/bugs/>. From R (Section 18.2.2) it is possible to access WinBUGS by using the R package WinBUGS or glmmBUGS.

References

- Abayomi, K., Gelman, A., and Levy, M. (2008) ‘Diagnostics for multivariate imputations’. *Applied Statistics*, 57, 273–291.
- Achen, C.H. (2005) ‘Two-step hierarchical estimation: Beyond regression analysis’. *Political Analysis*, 13, 447–456.
- Agresti, A. (2002) *Categorical Data Analysis*, 2nd edn. New York: Wiley.
- Agresti, A., and Natarajan, R. (2001) ‘Modeling clustered ordered categorical data’. *International Statistical Review*, 69, 345–371.
- Aitkin, M., Anderson, D., and Hinde, J. (1981) ‘Statistical modelling of data on teaching styles’. *Journal of the Royal Statistical Society, Series A*, 144, 419–461.
- Aitkin, M., and Longford, N. (1986) ‘Statistical modelling issues in school effectiveness studies’ (with discussion). *Journal of the Royal Statistical Society, Series A*, 149, 1–43.
- Alker, H.R. (1969) ‘A typology of ecological fallacies’. In: M. Dogan and S. Rokkan (eds), *Quantitative Ecological Analysis in the Social Sciences*, pp. 69–86. Cambridge, MA: MIT Press.
- Alonso, A., Litière, S., and Laenen, A. (2010) ‘A note on the indeterminacy of the random-effects distribution in hierarchical models’. *The American Statistician*, 64, 318–324.
- Anderson, D., and Aitkin, M. (1985) ‘Variance components models with binary response: Interviewer variability’. *Journal of the Royal Statistical Society, Series B*, 47, 203–210.
- Atkinson, A.C. (1985) *Plots, Transformations, and Regression*. Oxford: Clarendon Press.
- Asparouhov, T. (2006) ‘General multi-level modeling with sampling weights’. *Communications in Statistics – Theory and Methods*, 35, 439–460.
- Austin, P.C. (2010) ‘Estimating multilevel logistic regression models when the number of clusters is low: A comparison of different statistical software procedures’. *International Journal of Biostatistics*, 6, Article 16.
- Baltagi, B.H. (2008) *Econometric Analysis of Panel Data*, 4th edn. Chichester: Wiley.
- Bates, D.M. (2010) *lme4: Mixed-effects modeling with R*. In preparation. <http://lme4.r-forge.r-project.org/book/>
- Bauer, D.J. (2009) ‘A note on comparing the estimates of models for cluster-correlated or longitudinal data with binary or ordinal outcomes’. *Psychometrika* 74, 97–105.
- Bauer, D.J., and Curran, P.J. (2005) ‘Probing interactions in fixed and multilevel regression: Inferential and graphical techniques’. *Multivariate Behavioral Research*, 40, 373–400.
- Bell, R.M., and McCaffrey, D.F. (2002) ‘Bias reduction in standard errors for linear regression with multi-stage samples’. *Survey Methodology*, 28(2), 169–179.
- Beretvas, S.N. (2011) ‘Cross-classified and multiple-membership models’. In: J.J. Hox and J.K. Roberts (eds), *Handbook of Advanced Multilevel Analysis*, pp. 313–334. New York: Routledge.
- Berk, R., and MacDonald, J. (2008) ‘Overdispersion and Poisson regression’. *Journal of Quantitative Criminology*, 24, 269–284.
- Berkhof, J., and Kampen, J.K. (2004) ‘Asymptotic effect of misspecification in the random part of the multilevel model’. *Journal of Educational and Behavioral Statistics*, 29, 201–218.

- Berkhof, J., and Snijders, T.A.B. (2001) 'Variance component testing in multilevel models'. *Journal of Educational and Behavioral Statistics*, 26, 133–152.
- Berridge, D.M., and Crouchley, R. (2011) *Multivariate Generalized Linear Mixed Models Using R*. London: Psychology Press, Taylor and Francis.
- Bertolet, M. (2008) To weight or not to weight? Incorporating sampling designs into model-based analyses. PhD dissertation, Carnegie Mellon University. <http://gradworks.umi.com/3326665.pdf>
- Bertolet, M. (2010) *To Weight or Not To Weight? A survey sampling simulation study*. Submitted for publication.
- Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975) *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.
- Blakely, T.A., and Woodward, A.J. (2000) 'Ecological effects in multi-level studies'. *Journal of Epidemiology and Community Health* 54, 367–374.
- Bliese, P.D. (2009) *Multilevel Modeling in R* (2.3). http://cran.r-project.org/doc/contrib/Bliese_Multilevel.pdf
- Bohrenstein, M., Hedges, L.V., Higgins, J.P.T., and Rothstein, H.R. (2009) *Introduction to Meta-Analysis*. New York: Wiley.
- Bosker, R.J., Snijders, T.A.B., and Guldemon, H. (2003) *PinT (Power in Two-level designs). Estimating Standard Errors of Regression Coefficients in Hierarchical Linear Models for Power Calculations. User's Manual Version 2.1*. Groningen: University of Groningen. <http://www.stats.ox.ac.uk/~snijders/multilevel.htm#progPINT>
- Box, G.E.P., and Cox, D.R. (1964) 'An analysis of transformations' (with discussion). *Journal of the Royal Statistical Society, Series B*, 26, 211–252.
- Box, G.E.P., Hunter, W.G., and Hunter, J.S. (1978) *Statistics for Experimenters*. New York: Wiley.
- Brandsma, H.P. and Knuver, J.W.M. (1989) 'Effects of school and classroom characteristics on pupil progress in language and arithmetic'. *International Journal of Educational Research*, 13, 777–788.
- Breslow, N.E., and Clayton, D.G. (1993) 'Approximate inference in generalized linear mixed models'. *Journal of the American Statistical Association*, 88, 9–25.
- Browne, W.J. (2004) 'An illustration of the use of reparameterisation methods for improving MCMC efficiency in crossed random effect models'. *Multilevel Modelling Newsletter*, 16(1), 13–25.
- Browne, W.J. (2009) *MCMC Estimation in MLwiN* (Version 2.13). Bristol: Centre for Multilevel Modelling, University of Bristol.
- Browne W.J., and Draper, D. (2000) 'Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models'. *Computational Statistics*, 15, 391–420.
- Browne, W.J., and Draper, D. (2006) 'A comparison of Bayesian and likelihood methods for fitting multilevel models' (with discussion). *Bayesian Analysis*, 1, 473–550.
- Browne, W.J., Goldstein, H. and Rasbash, J. (2001) 'Multiple membership multiple classification (MMMC) models'. *Statistical Modelling*, 1, 103–124.
- Browne, W.J., Lahi, M.G., and Parker, R.M.A. (2009) *A Guide to Sample Size Calculations for Random Effects Models via Simulation and the MLPowSim Software Package*. Bristol: University of Bristol.
- Browne, W.J., McCleery, R.H., Sheldon, B.C., and Pettifor, R.A. (2007) 'Using cross-classified multivariate mixed response models with application to life history traits in great tits (*Parus major*)'. *Statistical Modelling*, 7, 217–238.
- Bryk, A.S., Raudenbush, S.W., and Congdon, R.T. (1996) *HLM. Hierarchical Linear and Non-linear Modeling with the HLM/2L and HLM/3L Programs*. Chicago: Scientific Software International.
- Burr, D., and Doss, H. (2005) 'A Bayesian semiparametric model for random-effects meta-analysis'. *Journal of the American Statistical Association*, 100, 242–251.

- Burstein, L., Linn, R.L., and Capell, F.J. (1978) 'Analyzing multilevel data in the presence of heterogeneous within-class regressions'. *Journal of Educational Statistics*, 3, 347–383.
- Cameron, A.C., Gelbach, J.B., and Miller, D.L. (2008) Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, 90, 414–427.
- Cameron, A.C., and Trivedi, P.K. (1998) *Regression Analysis of Count Data*. Cambridge: Cambridge University Press.
- Carle, A.C. (2009) 'Fitting multilevel models in complex survey data with design weights: Recommendations'. *BMC Medical Research Methodology*, 9:49.
- Carpenter, J.R., and Kenward, M.G. (2008) *Missing Data in Randomised Controlled Trials – A Practical Guide*. Birmingham: National Institute for Health Research, Publication RM03/JH17/MK. <http://www.hpa.ac.uk/nihrmethology/reports/1589.pdf>
- Chen, J., Zhang, D., and Davidian, M. (2002) 'A Monte Carlo EM algorithm for generalized linear mixed models with flexible random-effects distribution'. *Biostatistics*, 3, 347–360.
- Chow, G.C. (1984) 'Random and changing coefficient models'. In: Z. Griliches and M.D. Intriligator (eds), *Handbook of Econometrics, Volume 2*. Amsterdam: North-Holland.
- Cleveland, W.S. (1979) 'Robust locally weighted regression and smoothing scatterplots'. *Journal of the American Statistical Association*, 74, 829–836.
- Cochran, W.G. (1977) *Sampling Techniques*, 3d edn. New York: Wiley.
- Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences*. 2nd edn. Hillsdale, NJ: Erlbaum.
- Cohen, J. (1992) 'A power primer'. *Psychological Bulletin*, 112, 155–159.
- Cohen, M. (1998) 'Determining sample sizes for surveys with data analyzed by hierarchical linear models'. *Journal of Official Statistics*, 14, 267–275.
- Cohen, M.P. (2005) 'Sample size considerations for multilevel surveys'. *International Statistical Review*, 73, 279–287.
- Coleman, J.S. (1990) *Foundations of Social Theory*. Cambridge, MA: Harvard University Press.
- Commenges, D. and Jacqmin, H. (1994) 'The intraclass correlation coefficient: Distribution-free definition and test'. *Biometrics*, 50, 517–526.
- Commenges, D., Letenneur, L., Jacqmin, H., Moreau, Th., and Dartigues, J.-F. (1994) 'Test of homogeneity of binary data with explanatory variables'. *Biometrics*, 50, 613–620.
- Congdon, P.D. (2010) *Applied Bayesian Hierarchical Methods*. Boca Raton, FL: Chapman & Hall/CRC.
- Cook, R.D., and Weisberg, S. (1982) *Residuals and Influence in Regression*. New York and London: Chapman & Hall.
- Cook, R.D., and Weisberg, S. (1999) *Applied Regression Including Computing and Graphics*. New York: Wiley.
- Copt, S., and Victoria-Feser, M.-P. (2006) 'High-breakdown inference for mixed linear models'. *Journal of the American Statistical Association* 101, 292–300.
- Cornfield, J., and Tukey, J.W. (1956) 'Average values of mean squares in factorials', *Annals of Mathematical Statistics*, 27, 907–949.
- Cox, D.R. (1990) 'Role of models in statistical analysis'. *Statistical Science*, 5, 169–174.
- Croon, M.A., and van Veldhoven, M.J. (2007) 'Predicting group-level outcome variables from variables measured at the individual level: A latent variable multilevel model'. *Psychological Methods*, 12, 45–57.
- Davidian, M., and Giltinan, D.M. (1995) *Nonlinear Models for Repeated Measurement Data*. London: Chapman & Hall.
- Davis, J.A., Spaeth, J.L., and Huson, C. (1961) 'A technique for analyzing the effects of group composition'. *American Sociological Review*, 26, 215–225.
- de Leeuw, J., and Kreft, I. (1986) 'Random coefficient models for multilevel analysis'. *Journal of Educational Statistics*, 11 (1), 57–85.

- de Leeuw, J. and Meijer, E. (eds) (2008a) *Handbook of Multilevel Analysis*. New York: Springer.
- de Leeuw, J. and Meijer, E. (2008b) 'Introduction to multilevel analysis'. In: J. de Leeuw and E. Meijer (eds), *Handbook of Multilevel Analysis*, pp. 1–75. New York: Springer.
- de Weerth, C. (1998) Emotion-related behavior in infants. PhD thesis, University of Groningen, The Netherlands.
- de Weerth, C., and van Geert, P. (2002) 'A longitudinal study of basal cortisol in infants: Intra-individual variability, circadian rhythm and developmental trends'. *Infant Behavior and Development*, 25, 375–398.
- Dekkers, H.P.J.M., Bosker, R.J., and Driessen, G.W.J.M. (2000) 'Complex inequalities of educational opportunities. A large-scale longitudinal study on the relation between gender, SES, ethnicity, and school success'. *Educational Research and Evaluation*, 6, 59–82.
- Demidenko, E. (2004) *Mixed Models. Theory and Applications*. Hoboken, NJ: Wiley.
- Diez-Roux, A. (1998) 'Bringing context back into epidemiology: Variables and fallacies in multilevel analysis'. *American Journal of Public Health* 88, 216–222.
- Diez-Roux, A. (2000) 'Multilevel analysis in public health research.' *Annual Review of Public Health*, 21, 171–192.
- Diggle, P.J., Heagerty, P.K., Liang, K.Y., and Zeger, S.L. (2002) *Analysis of Longitudinal Data* (2nd edn). Oxford: Oxford University Press.
- Dogan, M., and Rokkan, S. (eds) (1969) *Quantitative Ecological Analysis in the Social Sciences*. Cambridge, MA: MIT Press.
- Donner, A. (1986) 'A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model'. *International Statistical Review*, 54, 67–82.
- Donner, A., and Klar, N. (2000) *Design and analysis of cluster randomization trials in health research*. London: Arnold.
- Donner, A., and Wells, G. (1986) 'A comparison of confidence interval methods for the intraclass correlation coefficient'. *Biometrics*, 42, 401–412.
- Doolaard, S. (1999) *Schools in change or schools in chains?* Enschede: Twente University Press.
- Doran, H., Bates, D., Bliese, P. and Dowling, M. (2007) 'Estimating the multilevel Rasch model: With the lme4 package'. *Journal of Statistical Software*, 20(2).
- Dorman, J.P. (2008) 'The effect of clustering on statistical tests: An illustration using classroom environment data', *Educational Psychology*, 28, 583–595.
- Draper, D. (2008) 'Bayesian multilevel analysis and MCMC'. In: J. de Leeuw and E. Meijer (eds), *Handbook of Multilevel Analysis*, pp. 77–139. New York: Springer.
- Dueck, A., and Lohr, S. (2005) 'Robust estimation of multivariate covariance components'. *Biometrics*, 61, 162–169.
- DuMouchel, W.H., and Duncan, G.J. (1983) 'Using sample survey weights in multiple regression analysis of stratified samples'. *Journal of the American Statistical Association*, 78, 535–543.
- Duncan, O.D., Curzort, R.P., and Duncan, R.P. (1961) *Statistical Geography: Problems in Analyzing Areal Data*. Glencoe, IL: Free Press.
- Eberly, L.E., and Thackeray, L.M. (2005) 'On Lange and Ryan's plotting techniques for diagnosing non-normality of random effects'. *Statistics and Probability Letters*, 75, 77–85.
- Efron, B., and Morris, C.N. (1975) 'Data analysis using Stein's estimator and its generalizations'. *Journal of the American Statistical Association*, 74, 311–319.
- Eicker, F. (1963) 'Asymptotic normality and consistency of the least squares estimator for families of linear regressions'. *Annals of Mathematical Statistics*, 34, 447–456.
- Eisenhart, C. (1947) 'The assumptions underlying the analysis of variance'. *Biometrics*, 3, 1–21.
- Enders, C.K. and Tofghi, D. (2007) 'Centering predictor variables in cross-sectional multilevel models: a new look at an old issue'. *Psychological Methods*, 12(2), 121–138.
- Fielding, A. (2004a) 'The role of the Hausman test and whether higher level effects should be treated as random or fixed'. *Multilevel Modelling Newsletter*, 16(2), 3–9.

- Fielding, A. (2004b) ‘Scaling for residual variance components of ordered category responses in generalised linear mixed multilevel models’. *Quality and Quantity* 38, 425–433.
- Firebaugh, G. (1978) ‘A rule for inferring individual-level relationships from aggregate data’. *American Sociological Review*, 43, 557–572.
- Fisher, R.A. (1924) ‘On a distribution yielding the error functions of several well-known statistics’. *Proceedings of the International Mathematical Congress, Toronto*, pp. 805–813.
- Fisher, R.A. (1958) *Statistical Methods for Research Workers*, 13th edn. London: Hafner Press.
- Fisher, R.A., and Yates, F. (1963) *Statistical Tables for Biological, Agricultural, and Medical Research*. Edinburgh: Oliver & Boyd.
- Fitzmaurice, G.M., Laird, N.M., and Ware, J.H. (2004) *Applied Longitudinal Analysis*. Hoboken, NJ: Wiley.
- Fong, Y., Rue, H., and Wakefield, J. (2010) ‘Bayesian inference for generalized linear mixed models’. *Biostatistics* 11, 397–412.
- Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, 2nd edn. Thousand Oaks, CA: Sage.
- Freedman, D.A. (2006) ‘On the so-called “Huber Sandwich estimator” and “robust standard errors”’. *The American Statistician*, 60, 299–302.
- Fuller, W.A. (2009) *Sampling Statistics*. Hoboken, NJ: Wiley.
- Gardiner, J.C., Luo, Z.H., and Roman, L.A. (2009) ‘Fixed effects, random effects and GEE: What are the differences?’. *Statistics in Medicine*, 28, 221–239.
- Gelman, A. (2007) ‘Struggles with survey weighting and regression modeling’. *Statistical Science*, 22, 153–164.
- Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2004) *Bayesian Data Analysis*, 2nd edn. Boca Raton, FL: Chapman & Hall/CRC.
- Gelman, A., and Hill, J. (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.
- Gibbons, R.D., and Bock, R.D. (1987) ‘Trends in correlated proportions’. *Psychometrika*, 52, 113–124.
- Gibbons, R.D., and Hedeker, D. (1994) ‘Application of random-effects probit regression models’. *Journal of Consulting and Clinical Psychology*, 62, 285–296.
- Gibbons, R.D., and Hedeker, D. (1997) ‘Random effects probit and logistic regression models for three-level data’. *Biometrics*, 53, 1527–1537.
- Gilks, W.R., Richardson, S., and Spiegelhalter, D.J. (1996) *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.
- Glass, G.V., and Stanley, J.C. (1970) *Statistical Methods in Education and Psychology*. Englewood Cliffs, NJ: Prentice Hall.
- Glidden, D.V., and Vittinghoff, E. (2004) ‘Modelling clustered survival data from multicentre clinical trials’. *Statistics in Medicine*, 23, 369–388.
- Goldstein, H. (1986) ‘Multilevel mixed linear model analysis using iterative generalized least squares’. *Biometrika*, 73, 43–56.
- Goldstein, H. (1991) ‘Nonlinear multilevel models with an application to discrete response data’. *Biometrika*, 78, 45–51.
- Goldstein, H. (2011) *Multilevel Statistical Models*. 4th edn. London: Edward Arnold.
- Goldstein, H., Carpenter, J., Kenward, M.G., and Levin, K.A. (2009) ‘Multilevel models with multivariate mixed response types’. *Statistical Modelling*, 9, 173–197.
- Goldstein, H., and Healy, M.J.R. (1995) ‘The graphical presentation of a collection of means’. *Journal of the Royal Statistical Society, Series A*, 158, 175–177.
- Goldstein, H., and Rasbash, J. (1996) ‘Improved approximations for multilevel models with binary responses’. *Journal of the Royal Statistical Society, Series A*, 159, 505–513.

- Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G., and Healy, M. (1998) *A user's guide to MLwiN*. London: Multilevel Models Project, Institute of Education, University of London.
- Gouriéroux, C., and Montfort, A. (1996) *Simulation-Based Econometric Methods*. Oxford: Oxford University Press.
- Graham, J.W. (2009) 'Missing data analysis: Making it work in the real world'. *Annual Review of Psychology*, 60, 549–576.
- Graham, J.W., Olchowski, A.E., and Gilreath, T.D. (2007) 'How many imputations are really needed? Some practical clarifications of multiple imputation theory'. *Prevention Science*, 8, 206–213.
- Greene, W. (2008) *Econometric Analysis*, 6th edn. Upper Saddle River, NJ: Prentice Hall.
- Grilli, L., and Pratesi, M. (2004) 'Weighted estimation in multi-level ordinal and binary models in the presence of informative sampling designs'. *Survey Methodology*, 30, 93–103.
- Guldemond, H. (1994) *Van de kikker en de vijver [About the frog and the pond]*. PhD thesis, University of Amsterdam.
- Hagle, T.M., and Mitchell II, G.E. (1992) 'Goodness-of-fit measures for probit and logit'. *American Journal of Political Science*, 36, 762–784.
- Haldane, J.B.S. (1940) 'The mean and variance of χ^2 , when used as a test of homogeneity, when expectations are small'. *Biometrika*, 31, 346–355.
- Hamaker, E.L., and Klugkist, I. (2011) 'Bayesian estimation of multilevel models'. In: J.J. Hox and J.K. Roberts (eds), *Handbook of Advanced Multilevel Analysis*, pp. 137–161. New York: Routledge.
- Hand, D., and Crowder, M. (1996) *Practical Longitudinal Data Analysis*. London: Chapman & Hall.
- Harker, R., and Tymms, P. (2004) 'The effects of student composition on school outcomes'. *School Effectiveness and School Improvement*, 15, 177–199.
- Hauser, R.M. (1970) 'Context and consext: A cautionary tale.' *American Journal of Sociology*, 75, 645–654.
- Hauser, R.M. (1974) 'Contextual analysis revisited'. *Sociological Methods and Research*, 2, 365–375.
- Hausman, J.A. (1978) 'Specification tests in econometrics'. *Econometrica*, 46, 1251–1271.
- Hausman, J.A., and Taylor, W.E. (1981) 'Panel data and unobservable individual effects'. *Econometrica*, 49, 1377–1398.
- Hays, W.L. (1988) *Statistics*. 4th edn. New York: Holt, Rinehart and Winston.
- Heck, R.H., Thomas, S.L., and Tabata, L.N. (2010) *Multilevel and Longitudinal Modeling with IBM SPSS*. New York: Routledge.
- Hedeker, D. (2003) 'A mixed-effects multinomial logistic regression model'. *Statistics in Medicine*, 22, 1433–1446.
- Hedeker, D. (2008) 'Multilevel models for ordinal and nominal variables'. In: J. de Leeuw and E. Meijer (eds), *Handbook of Multilevel Analysis*, pp. 237–274. New York: Springer.
- Hedeker, D., and Gibbons, R.D. (1994) 'A random effects ordinal regression model for multilevel analysis'. *Biometrics*, 50, 933–944.
- Hedeker, D., and 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., and 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., and Waternaux, C. (1999) 'Sample size estimation for longitudinal designs with attrition: Comparing time-related contrasts between two groups.' *Journal of Educational and Behavioral Statistics*, 24, 70–93.
- Hedeker, D., and Gibbons, R.D. (2006) *Longitudinal Data Analysis*. Hoboken, NJ: Wiley.
- Hedges, L.V. (1992) 'Meta-analysis'. *Journal of Educational Statistics*, 17, 279–296.

- Hedges, L.V., and Hedberg, E.C. (2007) 'Intraclass correlation values for planning group-randomized trials in education'. *Educational Evaluation and Policy Analysis*, 29, 60–87.
- Hedges, L.V., and Olkin, I. (1985) *Statistical Methods for Meta-analysis*. New York: Academic Press.
- Hilden-Minton, J.A. (1995) Multilevel diagnostics for mixed and hierarchical linear models. PhD dissertation, Department of Mathematics, University of California, Los Angeles.
- Hill, P.W., and Goldstein, H. (1998) 'Multilevel modeling of educational data with cross-classification and missing identification for units'. *Journal of Educational and Behavioral Statistics*, 23, 117–128.
- Hodges, J.S. (1998) 'Some algebra and geometry for hierarchical linear models, applied to diagnostics'. *Journal of the Royal Statistical Society, Series B*, 60, 497–536.
- Hogan, J., Roy, J. and Korkontzelou, C. (2004) 'Handling drop-out in longitudinal studies'. *Statistics in Medicine*, 23, 1455–1497.
- Hosmer, D.W., and Lemeshow, S. (2000) *Applied Logistic Regression*, 2nd edn. New York: Wiley.
- Hox, J.J. (2010) *Multilevel Analysis: Techniques and Applications*, 2nd edn. Mahwah, NJ: Erlbaum.
- Hsiao, C. (1995) 'Panel analysis for metric data'. In: G. Arminger, C.C. Clogg, and M.E. Sobel (eds), *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, pp. 361–400. New York: Plenum Press.
- Huber, P.J. (1967) 'The behavior of maximum likelihood estimates under non-standard conditions'. In: L. LeCam and J. Neyman (eds), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 221–233. Berkeley: University of California Press.
- Hüttner, H.J.M. (1981) 'Contextuele analyse' [Contextual analysis]. In: M. Albinski (ed.), *Onderzoeks typen in de Sociologie*, pp. 262–288. Assen: Van Gorcum.
- Hüttner, H.J.M., and van den Eeden, P. (1995) *The Multilevel Design. A Guide with an Annotated Bibliography, 1980–1993*. Westport, CT: Greenwood Press.
- Ibrahim, J.G., and Molenberghs, G. (2009) 'Missing data methods in longitudinal studies: a review'. *Test*, 18, 1–43.
- Ihaka, R., and Gentleman, R. (1996) 'R: A language for data analysis and graphics'. *Journal of Computational and Graphical Statistics*, 5, 299–314.
- Jackman, S. (2009) *Bayesian Analysis for the Social Sciences*. Chichester: Wiley.
- Jang, M.J., Lee, Y., Lawson, A.B., and Browne, W.J. (2007) 'A comparison of the hierarchical likelihood and Bayesian approaches to spatial epidemiological modelling'. *Environmetrics*, 18, 809–821.
- Jank, W. (2006) 'Implementing and diagnosing the stochastic approximation EM algorithm'. *Journal of Computational and Graphical Statistics*, 15, 1–27.
- Johnson, N.L., Kotz, S., and Balakrishnan, N. (1995) *Continuous Univariate Distributions, Volume 2*, 2nd edn. New York: Wiley.
- Kaplan, D., and Elliott, P.R. (1997) 'A didactic example of multilevel structural equation modeling applicable to the study of organizations'. *Structural Equation Modeling*, 4, 1–24.
- Kasim, R.M., and Raudenbush, S.W. (1998) 'Application of Gibbs sampling to nested variance components models with heterogeneous within-group variance'. *Journal of Educational and Behavioral Statistics*, 23, 93–116.
- Kelley, T.L. (1927) *The Interpretation of Educational Measurements*. New York: World Books.
- King, G. (1997) *A Solution to the Ecological Inference Problem. Reconstructing Individual Behavior from Aggregate Data*. Princeton, NJ: Princeton University Press.
- Kish, L. (1992) 'Weighting for unequal P_i '. *Journal of Official Statistics*, 8, 183–200.
- Knuver, A.W.M. and Brandsma, H.P. (1993) 'Cognitive and affective outcomes in school effectiveness research'. *School Effectiveness and School Improvement*, 4, 189–204.
- Korn, E.L., and Graubard, B.I. (2003) 'Estimating variance components by using survey data'. *Journal of the Royal Statistical Society, Series B*, 65, 175–190.

- Kovačević, M., Rong, H., and You, Y. (2006) ‘Bootstrapping for variance estimation in multi-level models fitted to survey data’. *ASA Proceedings of the Survey Research Methods Section*, pp. 3260–3269. <http://www.amstat.org/sections/srms/proceedings/y2006/Files/JSM2006-000286.pdf>
- Kreft, I.G.G., and de Leeuw, J. (1998) *Introducing Multilevel Modeling*. London: Sage Publications.
- Kreft, I.G.G., de Leeuw, J., and Aiken, L. (1995) ‘The effect of different forms of centering in hierarchical linear models’. *Multivariate Behavioral Research*, 30, 1–22.
- Kuk, A.Y.C. (1995) ‘Asymptotically unbiased estimation in generalized linear models with random effects’. *Journal of the Royal Statistical Society, Series B*, 57, 395–407.
- Kuyper, H., and Van der Werf, M.P.C. (2003) *VOCL'99-1: Technisch rapport*. Groningen: GION.
- LaHuis, D.M., and Ferguson, M.W. (2009) ‘The accuracy of significance tests for slope variance components in multilevel random coefficient models’. *Organizational Research Methods*, 12, 418–435.
- Laird, N. (1978) ‘Nonparametric maximum likelihood estimation of a mixture distribution’. *Journal of the American Statistical Association*, 73, 805–811.
- Laird, N.M., and Ware, J.H. (1982) ‘Random-effects models for longitudinal data’. *Biometrics*, 38, 963–974.
- Lange, N., and Ryan, L. (1989) ‘Assessing normality in random effects models’. *Annals of Statistics*, 17, 624–642.
- Langford, I.H., and Lewis, T. (1998) ‘Outliers in multilevel data’. *Journal of the Royal Statistical Society, Series A*, 161, 121–160.
- Lazarsfeld, P.F., and Menzel, H. (1971) ‘On the relation between individual and collective properties’. In: A. Etzioni (ed.), *A Sociological Reader on Complex Organizations*, p. 499–516. New York: Holt, Rinehart & Winston.
- Leckie, G. and Charlton, C. (2011) *runmlwin: Stata module for fitting multilevel models in the MLwiN software package*. Bristol: Centre for Multilevel Modelling, University of Bristol.
- Lehmann, E.L., and Romano, J.R. (2005) *Testing Statistical Hypotheses*, 3rd edn. New York: Springer.
- Lesaffre, E., and Verbeke, G. (1998) ‘Local influence in linear mixed models’. *Biometrics*, 54, 570–582.
- Liang, K.Y., and Zeger, S.L. (1986) ‘Longitudinal data analysis using generalized linear models’. *Biometrika*, 73, 13–22.
- Littell, R.C., Milliken, G.A., Stroup, W.W., Wolfinger, R.D., and Schabenberger, B. (2006) *SAS® System for Mixed Models*, 2nd edn. Cary, NC: SAS Institute.
- Little, R.J. (2004) ‘To model or not to model? Competing modes of inference for finite population sampling’ *Journal of the American Statistical Association*, 99, 546–556.
- Little, R.J.A., and Rubin, D.B. (2002) *Statistical Analysis with Missing Data*, 2nd edn. Hoboken, NJ: Wiley.
- Lohr, S.L. (2010) *Sampling: Design and Analysis*, 2nd edn. Boston: Brooks/Cole.
- Long, J.S. (1997) *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage Publications.
- Longford, N.T. (1987) ‘A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects’. *Biometrika*, 74(4), 812–827.
- Longford, N.T. (1993) *Random Coefficient Models*. New York: Oxford University Press.
- Longford, N.T. (1994) ‘Logistic regression with random coefficients’. *Computational Statistics and Data Analysis*, 17, 1–15.
- Longford, N.L. (1995) ‘Random coefficient models’. In: G. Arminger, C.C. Clogg, and M.E. Sobel (eds), *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, pp. 519–577. New York: Plenum Press.

- Lord, F.M., and Novick, M.R. (1968) *Statistical Theory of Mental Test Scores*. Reading, MA: Addison-Wesley.
- Lukočienė, O., and Vermunt, J.K. (2007) 'A Comparison of multilevel logistic regression models with parametric and nonparametric random intercepts'. Submitted for publication.
- Lukočienė, O., and Vermunt, J.K. (2009) 'Logistic regression analysis with multidimensional random effects: A comparison of three approaches'. Submitted for publication.
- Lüdtke, O., Marsh, H.W., Robitzsch, A., Trautwein, U., Asparouhov, T., and Muthén, B. (2008) 'The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies'. *Psychological Methods*, 13, 203–229.
- Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000) 'WinBUGS – a Bayesian modelling framework: Concepts, structure, and extensibility'. *Statistics and Computing*, 10, 325–337.
- Maas, C.J.M., and Hox, J.J. (2004) 'The influence of violations of assumptions on multilevel parameter estimates and their standard errors'. *Computational Statistics & Data Analysis*, 46, 427–440.
- Maas, C.J.M., and Hox, J.J. (2005) 'Sufficient sample sizes for multilevel modeling'. *Methodology*, 1, 86–92.
- Maas, C.J.M., and Snijders, T.A.B. (2003) 'The multilevel approach to repeated measures for complete and incomplete data'. *Quality and Quantity*, 37, 71–89.
- MacKinnon, J.G., and White, H. (1985) 'Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties'. *Journal of Econometrics*, 29, 305–325.
- Maddala, G.S. (1971) 'The use of variance component models in pooling cross section and time series data'. *Econometrica*, 39, 341–358.
- Maindonald, J., and Braun, J. (2007) *Data Analysis and Graphics Using R*, 2nd edn. Cambridge: Cambridge University Press.
- Mancl, L.A., and DeRouen, T.A. (2001) 'A covariance estimator for GEE with improved small-sample properties'. *Biometrics*, 57, 126–134.
- 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.
- Mason, A., Richardson, S., Plewis, I. and Best, N. (2010) Strategy for modelling non-random missing data mechanisms in observational studies using Bayesian methods. The BIAS project, London.
- Mason, W.M., Wong, G.M., and Entwistle, B. (1983) 'Contextual analysis through the multilevel linear model'. In: S. Leinhardt (ed.), *Sociological Methodology – 1983–1984*, pp. 72–103. San Francisco: Jossey-Bass.
- Maxwell, S.E., and Delaney, H.D. (2004) *Designing Experiments and Analyzing Data*, 2nd edn. Mahwah, NJ: Lawrence Erlbaum Associated.
- McCullagh, P., and Nelder, J.A. (1989) *Generalized Linear Models*. 2nd edn. London: Chapman & Hall.
- McCulloch, C.E. (1997) 'Maximum likelihood algorithms for generalized linear mixed models'. *Journal of the American Statistical Association*, 92, 162–170.
- McCulloch, C.E., and Searle, S.R. (2001) *Generalized, Linear, and Mixed Models*. New York: Wiley.
- McKelvey, R.D., and Zavoina, W. (1975) 'A statistical model for the analysis of ordinal level dependent variables'. *Journal of Mathematical Sociology*, 4, 103–120.
- Mealli, F., and Rampichini, C. (1999) 'Estimating binary multilevel models through indirect inference'. *Computational Statistics and Data Analysis*, 29, 313–324.
- Miller, J.J. (1977) 'Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance'. *Annals of Statistics*, 5, 746–762.
- Mills, M. (2011) *Introducing Survival and Event History Analysis*. Thousand Oaks, CA, and London: Sage.
- Moerbeek, M. (2004) 'The consequence of ignoring a level of nesting in multilevel analysis'. *Multivariate Behavioral Research*, 39, 129–149.

- Moerbeek, M. (2005) 'Randomization of clusters versus randomization of persons within clusters: which is preferable?'. *The American Statistician* 59, 173–179.
- Moerbeek, M.N., and Teerenstra, S. (2011) 'Optimal design in multilevel experiments'. In: J.J. Hox and J.K. Roberts (eds), *Handbook of Advanced Multilevel Analysis*, pp. 257–281. New York: Routledge.
- Moerbeek, M., van Breukelen, G.J.P., and Berger, M.P.F. (2003) A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies. *Journal of Clinical Epidemiology*, 56, 341–350.
- Mok, M. (1995) 'Sample size requirements for 2-level designs in educational research'. *Multilevel Modeling Newsletter*, 7 (2), 11–15.
- Molenberghs, G., and Kenward, M.G. (2007) *Missing Data in Clinical Studies*. Chichester: Wiley.
- Molenberghs, G., and Verbeke, G. (2006) *Models for Discrete Longitudinal Data*. New York: Springer.
- Molenberghs, G., and Verbeke, G. (2007) 'Likelihood ratio, score, and Wald tests in a constrained parameter space'. *The American Statistician* 61, 22–27.
- Mosteller, F., and Tukey, J.W. (1977) *Data Analysis and Regression*. Reading, MA: Addison-Wesley.
- Muthén, B.O. (1994) 'Multilevel covariance structure analysis'. *Sociological Methods & Research*, 22, 376–398.
- Muthén, B., and Asparouhov, T. (2009) 'Multilevel regression mixture analysis'. *Journal of the Royal Statistical Society, Series A*, 172, 639–657.
- Muthén, L.K. and Muthén, B.O. (2010) *Mplus User's Guide*, 6th edn. Los Angeles: Muthén and Muthén.
- Neuhaus, J.M. (1992) 'Statistical methods for longitudinal and cluster designs with binary responses'. *Statistical Methods in Medical Research*, 1, 249–273.
- Neuhaus, J.M., and Kalbfleisch, J.D. (1998) 'Between- and within-cluster covariate effects in the analysis of clustered data'. *Biometrics*, 54, 638–645.
- Neuhaus, J.M., Kalbfleisch, J.D., and Hauck, W.W. (1991) 'A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data'. *International Statistical Review*, 59, 25–35.
- Ng, E.S.W., Carpenter, J.R., Goldstein, H., and Rasbash, J. (2006) 'Estimation in generalised linear mixed models with binary outcomes by simulated maximum likelihood'. *Statistical Modelling*, 6, 23–42.
- Nordberg, L. (1989) 'Generalized linear modeling of sample survey data'. *Journal of Official Statistics*, 5, 223–239.
- O'Campo, P. (2003) 'Invited commentary: Advancing theory and methods for multilevel models of residential neighborhoods and health'. *American Journal of Epidemiology*, 157, 9–13.
- OECD (2009) *PISA 2006 Technical Report*. <http://www.oecd.org/dataoecd/0/47/42025182.pdf>
- OECD (2010) *PISA 2009 Results: What Makes a School Successful? Resources, Policies and Practices (Volume IV)*. <http://dx.doi.org/10.1787/9789264091559-en>
- Opdenakker, M.C., and Van Damme, J. (1997) 'Centreren in multi-level analyse: implicaties van twee centreringsmethoden voor het bestuderen van schooleffectiviteit'. *Tijdschrift voor Onderwijsresearch*, 22, 264–290.
- Pan, H., and Goldstein, H. (1998) 'Multi-level repeated measures growth modelling using extended spline functions'. *Statistics in Medicine*, 17, 2755–2770.
- Pan, W., and Wall, M. (2002) 'Small-sample adjustments in using the sandwich variance estimator in generalized estimating equation'. *Statistics in Medicine*, 21, 1429–1441.
- Pedhazur, E.J. (1982) *Multiple Regression in Behavioral Research*. 2nd edn. New York: Holt, Rinehart and Winston.
- Peter, S., and Drobnić, S. (2010) 'Women and their memberships: A multilevel cross-national study on gender differentials in associational involvement'. Submitted for publication.

- Pfeffermann, D. (1993) 'The role of sampling weights when modeling survey data'. *International Statistical Review*, 61, 317–337.
- Pfeffermann, D., Moura, F.A.D.S., and Silva, P.L.D.N. (2006) 'Multi-level modelling under informative sampling'. *Biometrika*, 93, 943–959.
- Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., and Rasbash, J. (1998) 'Weighting for unequal selection probabilities in multi-level models'. *Journal of the Royal Statistical Society, Series B*, 60, 23–56.
- Pinheiro, J.C., and Bates, D.M. (1995) 'Approximations to the log-likelihood function in nonlinear mixed-effects models'. *Journal of Computational and Graphical Statistics*, 4, 12–35.
- Pinheiro, J.C., and Bates, D.M. (2000) *Mixed-Effects Models in S and S-Plus*. New York: Springer.
- Pinheiro, J.C., Liu, C., and Wu, Y.N. (2001) 'Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate *t*-distribution'. *Journal of Computational and Graphical Statistics*, 10, 249–276.
- Potthoff, R. F., Woodbury, M. A., and Manton, K. G. (1992) "Equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey weights under superpopulation models'. *Journal of the American Statistical Association*, 87, 383–396.
- Preacher, K.J., Curran, P.J., and 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.
- Pregibon, D. (1981) 'Logistic regression diagnostics'. *Annals of Statistics*, 9, 705–724.
- Rabe-Hesketh, S., and Skrondal, A. (2006) 'Multilevel modelling of complex survey data'. *Journal of the Royal Statistical Society, Series A*, 169, 805–827.
- Rabe-Hesketh, S., and Skrondal, A. (2008) *Multilevel and Longitudinal Modeling Using Stata*, 2nd edn. College Station, TX: Stata Press.
- Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2004) 'Generalized multilevel structural equation modelling'. *Psychometrika*, 69, 167–190.
- Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2005) 'Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects.' *Journal of Econometrics*, 128, 301–323.
- Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., and Solenberger, P. (2001) 'A multivariate technique for multiply imputing missing values using a sequence of regression models'. *Survey Methodology*, 27.1, 85–95.
- Rasbash, J. and Browne, W.J. (2001) 'Modeling non-hierarchical structures'. In: A.H. Leyland and H. Goldstein (eds), *Multilevel Modeling of Health Statistics*, pp. 93–105. Chichester: Wiley.
- Rasbash, J. and Browne, W.J. (2008) 'Non-hierarchical multilevel models'. In: J. de Leeuw and E. Meijer (eds), *Handbook of Multilevel Analysis*, pp. 301–334. New York: Springer.
- Rasbash, J., and Goldstein, H. (1994) 'Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model.' *Journal of Educational and Behavioral Statistics*, 19, 337–350.
- Rasbash, J., Steele, V., Browne, W.J., and Goldstein, H. (2009) *A User's Guide to MLwiN*. Bristol: Centre for Multilevel Modelling, University of Bristol.
- Rasbash, J., and Woodhouse, G. (1995) *MLn: Command Reference*. London: Multilevel Models Project, Institute of Education, University of London.
- Raudenbush, S.W. (1993) 'A crossed random effects model for unbalanced data with applications in cross sectional and longitudinal research'. *Journal of Educational Statistics*, 18, 321–349.
- Raudenbush, S.W. (1995) 'Hierarchical linear models to study the effects of social context on development'. In: Gottman, J.M. (ed.), *The Analysis of Change*, pp. 165–201. Mahwah, NJ: Lawrence Erlbaum Ass.
- Raudenbush, S.W. (1997) 'Statistical analysis and optimal design for cluster randomized trials'. *Psychological Methods*, 2, 173–185.

- Raudenbush, S.W., and Bryk, A.S. (1986) 'A hierarchical model for studying school effects'. *Sociology of Education*, 59, 1–17.
- Raudenbush, S.W., and Bryk, A.S. (1987) 'Examining correlates of diversity'. *Journal of Educational Statistics*, 12, 241–269.
- Raudenbush, S.W., and Bryk, A.S. (2002) *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd edn. Thousand Oaks, CA: Sage.
- Raudenbush, S.W., and Liu, X. (2000) 'Statistical power and optimal design for multisite randomized trials'. *Psychological Methods*, 5, 199–213.
- Raudenbush, S.W., and Liu, X. (2001) 'Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change'. *Psychological Methods*, 6, 387–401.
- Raudenbush, S.W., Martinez, A., and Spybrook, J. (2007) 'Strategies for improving precision in group-randomized experiments'. *Educational Evaluation and Policy Analysis*, 1, 5–29.
- Raudenbush, S.W., and Sampson, R. (1999a) 'Assessing direct and indirect effects in multilevel designs with latent variables'. *Sociological Methods and Research*, 28, 123–153.
- Raudenbush, S.W., and Sampson, R. (1999b) "Econometrics": Toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods'. *Sociological Methodology*, 29, 1–41.
- Raudenbush, S.W., Yang, M.-L., and 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.
- Rekers-Mombarg, L.T.M., Cole, L.T., Massa, G.G., and Wit, J.M. (1997) 'Longitudinal analysis of growth in children with idiopathic short stature.' *Annals of Human Biology*, 24, 569–583.
- Richardson, A.M. (1997) 'Bounded influence estimation in the mixed linear model'. *Journal of the American Statistical Association*, 92, 154–161.
- Robinson, W.S. (1950) 'Ecological correlations and the behavior of individuals'. *American Sociological Review*, 15, 351–357.
- Rodríguez, G. (2008) 'Multilevel generalized linear models'. In: J. de Leeuw and E. Meijer (eds), *Handbook of Multilevel Analysis*, pp. 335–376. New York: Springer.
- Rodríguez, G. and Goldman, N. (1995) 'An assessment of estimation procedures for multilevel models with binary responses'. *Journal of the Royal Statistical Society, Series A*, 158, 73–89.
- Rodríguez, G. and Goldman, N. (2001) 'Improved estimation procedures for multilevel models with 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*, rev. edn. Newbury Park, CA: Sage Publications.
- Rotondi, M.A. and Donner, A. (2009) 'Sample size estimation in cluster randomized educational trials: An empirical Bayes approach. *Journal of Educational and Behavioral Statistics*, 34, 229–237.
- Roy, J., and Lin, X. (2005) 'Missing covariates in longitudinal data with informative dropout: Bias analysis and inference'. *Biometrics*, 61, 837–846.
- Rubin, D.B. (1976) 'Inference and missing data'. *Biometrika*, 63, 581–592.
- Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
- Rue, H., Martino, S., and Chopin, N. (2009) 'Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations' (with discussion). *Journal of the Royal Statistical Society, Series B*, 71, 319–392.
- Ruiter, S., and van Tubergen, F. (2009) 'Religious attendance in cross-national perspective: A multilevel analysis of 60 countries'. *American Journal of Sociology*, 115, 863–895.
- Ryan, T.P. (1997) *Modern Regression Methods*. New York: Wiley.

- Sampson, R.J., Morenoff, J.D., and Gannon-Rowley, T. (2002) ‘Assessing “neighborhood effects”: social process and new directions in research’. *Annual Review of Sociology*, 28, 443–478.
- Schafer, J.L. (1997) *Analysis of Incomplete Multivariate Data*. New York: Chapman & Hall.
- Schafer, J.L., and Graham, J.W. (2002) ‘Missing data: Our view of the state of the art’. *Psychological Methods*, 7, 147–177.
- Schafer, J.L. and Yucel, R.M. (2002) ‘Computational strategies for multivariate linear mixed-effects models with missing values’. *Journal of Computational and Graphical Statistics*, 11, 437–457.
- Searle, S.R. (1956) ‘Matrix methods in components of variance and covariance analysis’. *Annals of Mathematical Statistics*, 29, 167–178.
- Searle, S.R., Casella, G., and McCulloch, C.E. (1992) *Variance Components*. New York: Wiley.
- Seber, G.A.F., and Wild, C.J. (1989) *Nonlinear Regression*. New York: Wiley.
- Self, G.S., and Liang, K.-Y. (1987) ‘Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions’. *Journal of the American Statistical Association*, 82, 605–610.
- Seltzer, M.H. (1993) ‘Sensitivity analysis for fixed effects in the hierarchical model: A Gibbs sampling approach’. *Journal of Educational Statistics*, 18, 207–235.
- Seltzer, M., and Choi, K. (2002) ‘Model checking and sensitivity analysis for multilevel models’. In: N. Duan and S. Reise (eds), *Multilevel Modeling: Methodological Advances, Issues, and Applications*. Hillsdale, NJ: Lawrence Erlbaum.
- Seltzer, M.H., Wong, W.H.H., and Bryk A.S. (1996) ‘Bayesian analysis in applications of hierarchical models: Issues and methods’. *Journal of Educational Statistics*, 21, 131–167.
- Shadish, W.R., Cook, T.D., and Campbell, D.T. (2002) *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.
- Shavelson, R.J., and Webb, N.M. (1991) *Generalizability Theory: A Primer*. Newbury Park, CA: Sage Publications.
- Siddiqui, O., Hedeker, D., Flay, B.R., and Hu, F. (1996) ‘Intraclass correlation estimates in a school-based smoking prevention study’. *American Journal of Epidemiology*, 144, 425–433.
- Singer, J.D. (1998a) ‘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. (1998b) ‘Fitting multilevel models using SAS PROC MIXED’. *Multilevel Modeling Newsletter*, 10(2), 5–9.
- Singer, J.D. (2002) ‘Fitting individual growth models using SAS PROC MIXED’. In: D.S. Moskowitz and S.L. Hershberger (eds), *Modeling intraindividual variability with repeated measures data*, pp. 135–170. Hillsdale, NJ: Lawrence Erlbaum.
- Singer, J.D., and Willett, J.B. (2003) *Applied Longitudinal Data Analysis*. New York: Oxford University Press.
- Skinner, C.J. (1989) ‘Domain means, regression and multivariate analysis’. In: C.J. Skinner, D. Holt, and T.M.F. Smith (eds), *Analysis of Complex Surveys*, pp. 59–87. New York: Wiley.
- Skrondal, A., and Rabe-Hesketh, S. (2004) *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton, FL: Chapman & Hall/CRC.
- Smith, T.M.F. (1994) ‘Sample surveys 1975–1990: An age of reconciliation?’ (with discussion). *International Statistical Review*, 62, 5–34.
- Snijders, J.Th., and Welten, V.J. (1968) *The I.S.I. School Achievement and Intelligence Tests, Form I and II*. Groningen: Wolters-Noordhoff.
- Snijders, T.A.B. (1996) ‘Analysis of longitudinal data using the hierarchical linear model’. *Quality & Quantity*, 30, 405–426.
- Snijders, T.A.B. (2001) ‘Sampling’. In: A. Leyland and H. Goldstein (eds), *Multilevel Modelling of Health Statistics*, pp. 159–174. Chichester: Wiley.

- Snijders, T.A.B. (2005) 'Power and sample size in multilevel linear models'. In: B.S. Everitt and D.C. Howell (eds), *Encyclopedia of Statistics in Behavioral Science. Volume 3*, pp. 1570–1573. Chichester: Wiley.
- Snijders, T.A.B., and Berkhof, J. (2008) 'Diagnostic checks for multilevel models'. In: J. de Leeuw and E. Meijer (eds), *Handbook of Multilevel Analysis*, pp. 141–175. New York: Springer.
- Snijders, T.A.B., and Bosker, R.J. (1993) 'Standard errors and sample sizes for two-level research'. *Journal of Educational Statistics*, 18, 237–259.
- Snijders, T.A.B., and Bosker, R.J. (1994) 'Modeled variance in two-level models'. *Sociological Methods & Research*, 22, 342–363.
- Spybrook, J., Raudenbush, S.W., Congdon, R. and Martinez, A. (2009) *Optimal Design for Longitudinal and Multilevel Research: Documentation for the "Optimal Design" Software, Version 2.0*. Ann Arbor, MI: University of Michigan.
- Stapleton, L. (2002) 'The incorporation of sample weights into multilevel structural equation models'. *Structural Equation Modeling*, 9, 475–502.
- StataCorp (2009) *Stata Statistical Software: Release 11*. College Station, TX: StataCorp.
- Staudenmayer, J., Lake, E.E., and Wand, M.P. (2009) 'Robustness for general design mixed models using the t-distribution'. *Statistical Modelling*, 9, 235–255.
- Steele, F. (2008) 'Multilevel models for longitudinal data'. *Journal of the Royal Statistical Society, Series A*, 171, 5–19.
- Sterba, S.K. (2009) 'Alternative model-based and design-based frameworks for inference from samples to populations: From polarization to integration'. *Multivariate Behavioral Research*, 44, 711–740.
- Stevens, J. (2009) *Applied Multivariate Statistics for the Social Sciences*, 5th edn. New York: Routledge.
- Stiratelli, R., Laird, N.M., and Ware, J.H. (1984) 'Random-effects models for serial observations with binary response'. *Biometrics*, 40, 961–971.
- Stram, D.O., and Lee, J.W. (1994) 'Variance components testing in the longitudinal mixed effects model'. *Biometrics*, 50, 1171–1177. Correction (1995) in vol. 51, p. 1196.
- Swamy, P.A.V.B. (1971) *Statistical Inference in Random Coefficient Regression Models*. New York: Springer.
- Tacq, J. (1986) *Van Multiniveau Probleem naar Multiveau Analyse*. Rotterdam: Department of Research Methods and Techniques, Erasmus University.
- Tanner, M.A., and Wong, W.H. (1987) 'The calculation of posterior distributions by data augmentation' (with discussion). *Journal of the American Statistical Association*, 82, 528–550.
- Thompson, L. A. (2009) 'R (and S-PLUS) manual to accompany Agresti's Categorical Data Analysis (2002), 2nd edition'.
- Timmermans, A.C., Snijders, T.A.B., and Bosker, R.J. (forthcoming) 'In search of value added in case of complex school effects'.
- Tranmer, M., and Steel, D.G. (2001) 'Ignoring a level in a multilevel model: Evidence from UK census data'. *Environment and Planning A*, 33, 941–948.
- Tuerlinckx, F., Rijmen, F., Verbeke, G., and De Boeck, P. (2006) 'Statistical inference in generalized linear mixed models: A review'. *British Journal of Mathematical and Statistical Psychology*, 59, 225–255.
- Twisk, J.W.R. (2006) *Applied Multilevel Analysis. A Practical Guide*. Cambridge: Cambridge University Press.
- van Buuren, S. (2007) 'Multiple imputation of discrete and continuous data by fully conditional specification'. *Statistical Methods in Medical Research*, 16, 219–242.
- van Buuren, S. (2011) 'Multiple imputation of multilevel data'. In: J.J. Hox and J.K. Roberts (eds), *Handbook of Advanced Multilevel Analysis*, pp. 173–196. New York: Routledge.

- van Buuren, S., Brand, J.P.L., Groothuis-Oudshoorn, C.G.M., and Rubin, D.B. (2006) 'Fully conditional specification in multivariate imputation'. *Journal of Statistical Computation and Simulation*, 76, 1049–1064.
- van den Noortgate, W., Opdenakker, M.-C., and Onghena, P. (2005) 'The effects of ignoring a level in multilevel analysis'. *School Effectiveness and School Improvement*, 16, 281–303.
- van der Leeden, R., Meijer, E., and Busing, F.M.T.A. (2008) 'Resampling multilevel models'. In: J. de Leeuw and E. Meijer (eds), *Handbook of Multilevel Analysis*, pp. 401–433. New York: Springer.
- van Mierlo, H., Vermunt, J.K., and Rutte, C.G. (2009) 'Composing group-level constructs from individual-level survey data'. *Organizational Research Methods*, 12, 368–392.
- van Yperen, N.W., and Snijders, T.A.B. (2000) 'Multilevel analysis of the demands-control model'. *Journal of Occupational Health Psychology*, 5, 182–190.
- Veall, M.R., and Zimmermann, K.F. (1992) 'Pseudo- R^2 's in the ordinal probit model'. *Journal of Mathematical Sociology*, 16, 333–342.
- Verbeke, G., and Lesaffre, E. (1996) 'A linear mixed-effects model with heterogeneity in the random-effects population'. *Journal of the American Statistical Association*, 91, 217–221.
- Verbeke, G., and 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., and Molenberghs, G. (1997) *Linear Mixed Models in Practice. A SAS-oriented Approach*. Lecture Notes in Statistics, 126. New York: Springer.
- Verbeke, G., and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data*. New York: Springer.
- Verbeke, G., and Molenberghs, G. (2003) 'The use of score tests for inference on variance components'. *Biometrics*, 59, 254–262.
- Vermeulen, C.J., and Bosker, R.J. (1992) *De Omvang en Gevolgen van Deeltijdsarbeid en Volledige Inzetbaarheid in het Basisonderwijs*. Enschede: University of Twente.
- Vermunt, J.K. (2003) 'Multilevel latent class models'. *Sociological Methodology*, 33, 213–239.
- Vermunt, J.K. (2008) 'Latent class and finite mixture models for multilevel data sets'. *Statistical Methods in Medical Research*, 17, 33–51.
- Vermunt, J.K. (2011) 'Mixture models for multilevel data sets'. In: J.J. Hox and J.K. Roberts (eds), *Handbook of Advanced Multilevel Analysis*, pp. 59–81. New York: Routledge.
- Vermunt, J.K. and Magidson, J. (2005a) *Latent GOLD 4.0 User's Guide*. Belmont, MA: Statistical Innovations Inc.
- Vermunt, J.K., and Magidson, J. (2005b) *Technical Guide for Latent GOLD 4.0: Basic and Advanced*. Belmont, MA: Statistical Innovations Inc.
- Waternaux, C., Laird, N.M., and Ware, J.H. (1989) 'Methods for analysis of longitudinal data: Blood lead concentrations and cognitive development.' *Journal of the American Statistical Association*, 84, 33–41.
- Wedel, M. and DeSarbo, W. (2002) 'Mixture regression models'. In: J. Hagenaars and A. McCutcheon (eds), *Applied Latent Class Analysis*, pp. 366–382. Cambridge: Cambridge University Press.
- West, B.T., Welch, K.B. and Gałecki, A.T. (2007) *Linear Mixed Models: A Practical Guide Using Statistical Software*. Boca Raton, FL: Chapman & Hall/CRC.
- White, H. (1980) 'A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity'. *Econometrica* 48, 817–830.
- Wilson, D.B., and Lipsey, M.W. (2001) *Practical Meta-analysis*. Thousand Oaks: Sage.
- Windmeijer, F.A.G. (1995) 'Goodness-of-fit measures in binary choice models'. *Econometric Reviews*, 14, 101–116.

- Windzio, M. (2006) 'The problem of time dependent explanatory variables at the context-level in discrete time multilevel event history analysis: A comparison of models considering mobility between local labour markets as an example'. *Quality and Quantity*, 40, 175–185.
- Winship, C., and Mare, R.D. (1984) 'Regression models with ordinal variables'. *American Sociological Review*, 49, 512–525.
- Wittek, R., and Wielers, R. (1998) 'Gossip in organizations'. *Computational and Mathematical Organization Theory*, 4, 189–204.
- Wright, D.B., and London, K. (2009) 'Multilevel modelling: Beyond the basic applications'. *British Journal of Mathematical and Statistical Psychology*, 62, 439–456.
- Wu, L. (2010) *Mixed Effects Models for Complex Data*. Boca Raton, FL: Chapman & Hall/CRC.
- Yau, K.K.W., and Kuk, A.Y.C. (2002) 'Robust estimation in generalized linear mixed models'. *Journal of the Royal Statistical Society, Series B*, 64, 101–117.
- Yuan, K.-H., and Bentler, P.M. (2002) 'On normal theory based inference for multilevel models with distributional violations'. *Psychometrika*, 67, 539–562.
- Yuan, Y., and Little, R.J.A. (2007) 'Parametric and semiparametric model-based estimates of the finite population mean for two-stage cluster samples with item nonresponse'. *Biometrics*, 63, 1172–1180.
- Yucel, R. (2008) 'Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response'. *Philosophical Transactions of the Royal Society, Series A*, 366, 2389–2403.
- Zaccarin, S., and Donati, C. (2009) 'The use of sampling weights in multilevel analysis of PISA data'. In: *Proceedings, First Italian Conference on Survey Methodology (ITACOSM09)* (Siena, June 10–12, 2009), pp. 161–164. Siena: Tipografia Sienese.
- Zeger, S.L., and Karim, M.R. (1991) 'Generalized linear models with random effects: A Gibbs sampling approach'. *Journal of the American Statistical Association*, 86, 79–86.
- Zeger S.L., Liang, K., and Albert, P. (1988) 'Models for longitudinal data: A generalised estimating equation approach'. *Biometrics*, 44, 1049–1060.
- Zhao, E., and Yucel, R.M. (2009) 'Performance of sequential imputation method in multilevel applications'. In *American Statistical Association Proceedings of the Survey Research Methods Section*. (American Statistical Association, Alexandria, VA), pp. 2800–2810.
- Zucker, D.M., Lieberman, O., and Manor, O. (2000) 'Improved small sample inference in the mixed linear model: Bartlett correction and adjusted likelihood'. *Journal of the Royal Statistical Society, Series B*, 62, 827–838.

Index

- adaptive quadrature, 300, 322
aggregation, 11, 14–16, 26, 33, 39, 55,
83, 102, 106, 154, 155, 157
of dichotomous data, 301
algorithms, 61, 89, 300–301, 322
alternative hypothesis, 192
analysis of covariance, 23, 30, 45, 47,
71, 100, 155
analysis of variance, 17, 22, 49, 100
analytic inference, 217, 222, 245
analytical weights, 220, 245
ANCOVA, *see* analysis of covariance
ANOVA, *see* analysis of variance
assumptions, 83, 107, 152, 153, 173
of hierarchical linear model, 75,
153–154
of random intercept model, 51, 55
autocorrelated residuals, 280, 281
- Bayesian information criterion (BIC),
202
Bayesian statistics, *see* empirical Bayes,
63, 137, 138, 144, 146,
194–197, 203, 244, 300, 322,
325, 331
between-group correlation, 32–35, 40,
285–288
between-group covariance matrix, 285
between-group regression, 1, 15, 27–31,
40, 56–60, 72, 83, 88–89, 95,
96, 102, 106, 145, 149, 151,
154, 173, 234, 258
in three-level model, 70
between-group variance, 18, 20, 21, 43
for dichotomous data, 292, 296
between-subjects variables, 251
BIC, *see* Bayesian information criterion
binary variables, *see* dichotomous
variables
Bonferroni correction, 170,
175, 278
bootstrap, 325
bootstrap standard errors, 244
Box–Cox transformation, 157, 174
budget constraint, 183, 186, 188,
189, 191
- categorical variable, 96, 97, 108, 124
ordered, *see* ordered categorical
variable
centering, 58, 81, 87–89, 93, 156
changing covariates, 276, 281
chi-squared test, 100, 291, 292, 321
Cholesky decomposition, 326
classification graph, 206, 211,
213, 215
cluster randomized trial, 184, 187, 193
cluster-robust standard error, *see*
sandwich estimator
clusters, 6, 7, 13, 24, 39, 42, 198–200,
223, 227, 228, 237
combination of estimates, 36, 40
combination of tests, 36, 40, 229
comparative standard error, *see* standard
error, comparative
complete data vector, 252, 281, 282
components of variance, *see* variance
components
compound symmetry model, 249, 256,
261, 280

- confidence interval, 65
for random part parameters, 100–101, 108
contextual analysis, 1–2
contextual effects, 56, 154, 173
contextual variable, 155
contingency table, 291
Cook's distance, 168, 175
correlates of diversity, 127, 129
correlation ratio, 31
cost function, 182, 193
count data, 289, 314–320, 322
covariance matrix, 76, 90, 252–259, 281
cross-classification, 206, 215
cross-classified models, 205
cross-level inference, 17
cross-level interaction, 11, 13, 16, 43, 55, 72, 81–83, 92, 95, 103, 104, 106, 116, 154, 185, 251, 275, 281
cross-validation, 126, 154, 163
crossed random effects, 206–210, 215

data augmentation, 136, 138
degrees of freedom, 60, 94, 97, 108
deletion diagnostics, 167, 169, 170, 175
deletion standardized multivariate residual, 175
delta method, 100, 327
dependent variable, 9, 27, 42, 71
descriptive inference, 217, 245
design effect, 17, 23–24, 39, 179, 187, 192, 221, 224, 246
design of multilevel studies, 7, 133, 150, 176–193
design variables, 219, 222, 223, 226, 228, 234, 237, 238, 245, 246
design-based inference, 2, 218, 222, 224, 245
deviance, 97
deviance test, 50, 60, 89, 96–99, 108, 316
for multiple imputations, 141
deviation score, 58, 70, 95, 110, 154
diagnostic standard error, *see* standard error, diagnostic
dichotomous variables, 289–309, 320

disaggregation, 16–17, 39
discrete outcome variables, 228
discrete outcome variables(, 289
discrete outcome variables), 322
dummy variables, 96, 98, 120, 124
for longitudinal models, 249, 256
for multivariate models, 284, 288
duration analysis, 313–314

ecological fallacy, 1, 15, 17, 39, 59, 83, 258
effect size, 53, 109–114, 177–179, 192
effective sample size, *see* design effect, 221, 224, 245
EM algorithm, 61, 89
emergent proposition, 11, 13
empirical Bayes estimation, 62–63, 65, 73, 89, 93, 147, 196
empirical Bayes residuals, 165, 174
empty model, 17, 49–51, 62, 72
for dichotomous data, 291, 295–297, 320
for longitudinal data, 250, 263
multivariate, 257, 285
estimation methods, 60–61, 89–90, 300–301, 322
event history analysis, 263, 313–314, 321
exchangeability, 46, 47, 75
expected value, 5
explained variance, 109–114, 118, 156, 173
at level one, 112
for discrete dependent variable, 305, 311, 321
in longitudinal models, 261–262, 281
in random slope models, 113, 118
in three-level models, 113, 118
explanatory variables, 42, 51, 54, 71, 74, 82, 87, 152, 250, 264, 293, 320

F-test, 22, 23, 100, 108
Fisher scoring, 61, 89, 156, 168
Fisher's *Z* transformation, 229

- Fisher's combination of p -values, 36, 229
- fixed coefficients, 46, 71
- fixed effects models, 46–48, 71
- fixed occasion designs, 248–262, 280
- fixed or random effects, 44–48
- fixed part, 55, 57, 75, 87, 153, 156–158
- frequency data, 314, 322
- frequency weights, 220, 245
- frequentist statistics, 195, 203
- full information maximum likelihood, 134, 150, 258
- fully multivariate model, 256, 261, 281, 282
- generalizability theory, 25
- generalized estimating equations, 198, 204
- generalized linear models, 290
- Gibbs sampling, 138, 173, 331
- gllamm, 202, 300, 315
- goodness of fit, 174
- Greek letters, 5
- group size, 56, 61, 188, 189
- groups, 17, 42, 74
- Hausman specification test, 57, 95
- heteroscedasticity, 37, 75–76, 92, 119–129, 153, 173, 197, 221, 228, 244, 253, 258, 290, 320, 324, 325
- at level one, 119–129, 159–161, 174, 279
 - at level two, 128–129
 - test of level-one, 159–161
- hierarchical generalized linear models, 238, 289–320, 328, 331
- hierarchical linear model, 2, 3, 41, 49, 72, 74–93, 116
- assumptions, 75, 153–154
- hierarchical model, 102
- hierarchies
- imperfect, 205–215
- history of multilevel analysis, 1–2
- HLM, 37, 104, 207, 238, 300, 301, 311, 315, 324–325
- homoscedasticity, 43, 71, 119, 153, 159, 163, 173, 174, 197, 221, 232, 233
- IGLS, *see* iterated generalized least squares
- imputation, 135–151
- chained equations, 144–148, 150
- inclusion probabilities, 216, 245–246
- independent variable, 42, *see* explanatory variables, 71
- influence, 161, 172, 175
- influence diagnostics, 167–171, 175
- informative design, 222, 223, 225–231, 246
- interaction
- random, 75
- interactions, *see* cross-level interactions, 89, 102, 105, 129, 146, 151, 154, 155, 173
- intercept, 27, 43, 44, 71
- random, 46, 49
- intercept variance, 49, 51, 52, 72, 155, 190
- intercept-slope covariance, 77, 79, 103
- intercepts as outcomes, 1, 80
- intraclass correlation, 17–35, 39, 50, 52, 72, 100, 104, 155, 181, 188, 192
- for binary data, 291, 304–306, 321
 - for ordered categorical data, 311
 - for three-level model, 68
 - residual, *see* residual intraclass correlation
- iterated generalized least squares, 61, 89, 156, 168
- Kelley estimator, 63
- Laplace approximation, 300, 322
- latent class models, 173, 201–204, 330
- Latent Gold, 202, 315, 330
- latent variables, 62, 75, 80
- level, 7, 8, 13, 155, 173
- omission of, 155
- level of randomization, 187

- level-one residuals, *see* residuals,
 level-one
 level-one unit, 8, 13, 17, 25, 42, 67,
 74, 182
 level-one variance, 190
 level-two random coefficients, 75,
 153, 174
 level-two residuals, *see* residuals,
 level-two
 level-two unit, 8, 13, 17, 25, 42,
 67, 74, 182
 leverage, 169–171, 174
 likelihood ratio test, *see* deviance test
 linear model, 71
 link function, 294, 295, 300, 304, 314,
 315, 320, 321
 LISREL, 238
 listwise deletion, 130
 loess, *see* lowess smoother
 log-odds, 293–298, 320
 logistic distribution, 304
 logistic function, 296, 304, 320
 logistic regression, 133, 145, 149, 161,
 290, 294, 297
 logit function, 293–296, 320
 longitudinal data, 2, 9, 247–282
 lowess smoother, 162, 166, 174
 macro–micro relations, 10, 11, 13, 15
 macro-level propositions, 11, 13
 macro-level relations, 26
 macro-unit, *see* level-two unit
 MAR, *see* missingness at random
 marginal quasi-likelihood, 300, 322
 Markov chain Monte Carlo, 138,
 195–197, 203, 300, 322
 maximum likelihood, 22, 35, 50, 51, 60,
 72, 89, 97, 108, 156, 195,
 197–200, 202, 322
 MCAR, *see* missingness completely at
 random
 MCMC
 see Markov chain Monte Carlo, 138
 membership weights, 207, 211, 215
 meta-analysis, 36, 40
 random effects model for, 37
 micro–macro relations, 11, 13
 micro-level propositions, 10, 13, 15
 micro-level relations, 15, 26
 micro-unit, *see* level-one unit
 missing data, 56, 130–151, 248,
 257–259, 285
 listwise deletion, 130
 missingness at random,
 132–134, 150
 missingness completely at random,
 132–133, 149
 missingness indicators, 131–133, 149
 missingness not at random,
 132–133, 150
 misspecification, 152, 156, 159, 162,
 173, 198–200, 222
 mixed model, 1, 41–93
 MIXOR, 300, 311, 312, 325, 326
 MIXPREG, 315
 mixture models, 201–204
 ML, *see* maximum likelihood
 MLPowSim, 191, 325, 328, 330
 MLwiN, 37, 153, 191, 207, 238, 301,
 311, 312, 315, 325, 330
 MNAR, *see* missingness not at random
 model of interest, 134, 136, 139, 140,
 143, 144, 148–150, 222, 246
 model specification, 56–60, 76–77, 85,
 87–90, 93, 102–108, 152,
 154–175
 of three-level model, 90
 model-based inference, 2, 3, 48,
 217–219, 222–224, 244, 245
 Mplus, 202, 238, 330
 multilevel logistic regression,
 290–309, 320
 Multilevel Models Project, 2, 323
 multilevel negative binomial regression,
 318, 322
 multilevel ordered logistic regression
 model, 161, 310–313, 321
 multilevel ordered probit model,
 310, 311
 multilevel Poisson regression,
 314–320, 322
 multilevel probit regression,
 304, 305, 321

- multilevel proportional odds model, 310
multiple classification, 207, 213–215
multiple correlation coefficient, 109
multiple imputation, 328
multiple membership models, 205, 207, 210–215
multiple membership multiple classification, 207, 213–215
multiple regression, 43
multisite trial, 187, 193
multivariate analysis, 283
multivariate empty model, 285
multivariate multilevel model, 282–288
multivariate regression analysis, 260, 281
multivariate residual, 169, 175

negative binomial regression, 318, 322
nested data, 3, 6–9
nonlinear models, 268, 289–320
normal probability plot, 163, 164, 166, 174
notational conventions, 5
null hypothesis, 36, 94, 96–98, 191
numerical integration, 300, 322, 326

observed between-group correlation, 32
observed variance, 18–20
odds, 293
offset, 316
OLS, 1, 23, 27, 44, 46, 52, 54, 62, 71, 87, 100, 104, 111, 159, 162, 198, 228
one-sided test, 94, 98
one-step estimator, 156, 168
Optimal Design, 330
ordered categorical variable, 289, 310, 321
ordinary least squares, *see* OLS
outliers, 165, 174
overdispersion, 318–320, 322

panel data, 2, 247, 280
parallel test items, 180
parameter estimation, *see* estimation methods
penalized quasi-likelihood, 300, 322

permutation test, 292
piecewise linear function, 268, 281
PinT, 181, 329
plausible values, 136, 325
Poisson distribution, 314, 322
Poisson regression, 290, 314
polynomial function, 265–267, 281
polynomial random part, 265
polynomial trend analysis, 253
population, 17, 46, 48–49, 217
population between-group variance, 18
population of curves, 263, 281
population within-group variance, 18
posterior confidence intervals, 64–67, 73
posterior distribution, 195, 203
posterior intercept, 63
posterior mean, 62–67, 73, 147, 165, 174, 180, 195, 203
posterior slope, 89, 93
posterior standard deviation, 65, 73, 195
power, 82, 103, 106, 132, 177–179, 192–193
power analysis, 193, 325, 328–330
precision weights, 220, 245
predictor variable, 42, 69, 71
prior distribution, 195, 203
probability model, 2–3, 217, 218, 237
probability weighted estimate, 220, 245
probit regression, 304
profile likelihood, 100
pseudo-cluster randomized trial, 187, 193
pseudo-maximum likelihood, 238–243, 246
psychological test theory, 25, 63

R, 37, 138, 153, 191, 207, 300, 301, 315, 328, 330
 R^2 , *see* explained variance
random coefficient model, 1
random coefficient models, 75
random coefficients, 2, 41, 46, 47, 71, 75
random effects, 1, 2, 17, 18, 47, 49
comparison of intercepts, 66
discrete, 201–203
random effects ANOVA, 49

random intercept, 46, 49, 54, 72, 74
test, 97–98, 108
random intercept model, 41–73, 114,
249, 280
comparison with OLS model, 54
for dichotomous data,
297–299, 320
multivariate, 283–288
three-level, 67, 284
random part, 55, 75, 87, 153, 158–161
test, 97–101, 108
random slope, 75, 77, 82, 92, 114,
156, 174
explanation, 80–85, 92
test, 97–101, 106, 108, 156
random slope model, 74–93, 116
for dichotomous outcome variable,
302, 320
for longitudinal data, 253, 280
multivariate, 288
random slope variance, 75, 78
interpretation, 77
REALCOM, 138, 330
reliability, 25–26, 39, 62, 180
REML, *see* residual maximum
likelihood
repeated measures, 180, 247, 280
residual intraclass correlation, 52, 61,
75, 187, 252
residual iterated generalized least
squares, 61, 89
residual maximum likelihood, 22, 35,
51, 60, 72, 89, 97, 108, 257
residual variance, 51, 76, 80, 153, 160
for dichotomous data, 291
nonconstant, 120, 123, 127
residuals, 27, 43, 46, 48, 51, 71, 75,
80, 83
level-one, 153, 161–165, 170,
174, 175
level-two, 165–171, 174
multivariate, *see* multivariate
residual
non-normal distributions, 199
nonnormal distributions, 172

restricted maximum likelihood, *see*
residual maximum likelihood
RIGLS, *see* residual iterated generalized
least squares
robust estimators, 175
robust standard errors, 197–200, 204

sample
cluster, 216, 231, 245
multistage, 6, 7, 13, 177
simple random, 6
stratified, 216, 231, 245
two-stage, 7, 17, 23, 39, 179, 180,
183, 192, 216, 231, 244
sample size, 23, 176–193
for estimating fixed effects,
180–187
for estimating group mean, 180
for estimating intraclass
correlation, 188–190
for estimating population mean,
179–180
for estimating variance parameter,
190–191
sample surveys, 216–246
sampling designs, 7
sampling probabilities, 216, 245–246
sampling weights, 216–246
scaling of, 225, 246
sandwich estimator, 173, 175, 197–200,
204, 220, 238, 246
SAS, 260, 300, 301, 312, 315, 327
Satterthwaite approximation, 95
shift of meaning, 15, 39, 59
shrinkage, 63–64, 66, 67, 73, 89
significance level, 177, 192
simple random sample, 216, 219, 224,
232, 245
slope variance, 92
slopes as outcomes, 1, 80, 92
software, 153, 300, 315, 323–331
spline function, 121, 129, 158, 174, 253,
270, 281
cubic, 270, 272
quadratic, 271
SPSS, 260, 329

- standard error, 23, 37, 141,
178–179, 181
comparative, 65, 73
diagnostic, 65, 73, 165
of empirical Bayes estimate, 63, 65
of fixed coefficients, 155, 181
of intercept variance, 191
of intraclass correlation, 21, 188
of level-one variance, 191
of population mean, 24, 179
of posterior mean, 63, 65
of random part parameters, 90, 100,
108, 327
of standard deviation, 90, 100, 327
of variance, 90, 100, 327
standardized coefficients, 53
standardized multivariate residual, *see*
multivariate residual
standardized OLS residuals, 174
Stata, 238, 300, 311, 315, 328–329
success probability, 290, 293, 295,
298, 314
SuperMix, 300, 315, 326
superpopulation model, 3, 236, 246
survey, 216
survey weights, *see* sampling weights
- t*-ratio, 59, 94, 108
t-test, 59, 94, 108
for random slope, 156
paired samples, 257, 280
test, 94–101, 163, 177
for random intercept, *see* random
intercept
for random slope, *see* random slope
test of heterogeneity of proportions,
292, 321
textbooks, 4
three-level model, 67–71, 73, 90–93,
113, 282, 284, 321
for multivariate multilevel model,
282–288
threshold model, 304, 308, 310, 321
- total variance, 18
total-group correlation, 32–35
total-group regression, 27–31
transformation
for count data, 315, 322
of dependent variable, 161
of explanatory variables, 157, 173
true score, 25, 63, 80, 180
two-stage sample, *see* sample, two-stage
type I error, 178, 192, 196, 198
type II error, 178, 192
- unexplained variation, 41, 46, 48, 51,
71, 75, 80, 111, 306
- V-known problem, 38, 324
variable occasion designs, 263–279, 281
variance components, 18, 109,
114–118, 188
- Wald test, 94, 96, 108, 141
weighted least squares, 220–222, 245
weights, *see* membership weights, *see*
sampling weights, *see*
precision weights
WinBUGS, 202, 331
within-group centering, 58, 87, 156
within-group correlation, 32–35, 40,
285, 286, 288
within-group covariance matrix, 285
within-group deviation score, *see*
deviation score
within-group regression, 1, 15, 27–31,
40, 56–60, 72, 83, 88–89, 95,
96, 102, 106, 145, 149, 151,
154, 173, 234, 258
in three-level model, 70
within-group variance, 18, 21
for dichotomous data, 293
within-subjects design, 249
within-subjects variable, 251
- zero variance estimate, 21, 61, 84, 85, 89

