Multilevel and Longitudinal Modeling Using Stata

Volume I: Continuous Responses

Fourth Edition

SOPHIA RABE-HESKETH
University of California–Berkeley

ANDERS SKRONDAL
Norwegian Institute of Public Health
University of Oslo
University of California–Berkeley

A Stata Press Publication
StataCorp LLC
College Station, Texas

 First edition 2005
Second edition 2008
Third edition 2012
Fourth edition 2022

Published by Stata Press, 4905 Lakeway Drive, College Station, Texas 77845
Typeset in LaTeX 2_ε
Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

Print ISBN-10: 1-59718-136-6 (volumes I and II)
Print ISBN-10: 1-59718-137-4 (volume I)
Print ISBN-10: 1-59718-138-2 (volume II)
Print ISBN-13: 978-1-59718-136-5 (volumes I and II)
Print ISBN-13: 978-1-59718-137-2 (volume I)
Print ISBN-13: 978-1-59718-138-9 (volume II)
ePub ISBN-10: 1-59718-309-1 (volumes I and II)
ePub ISBN-10: 1-59718-310-5 (volumes I)
ePub ISBN-10: 1-59718-311-3 (volumes II)
ePub ISBN-13: 978-1-59718-309-3 (volumes I and II)
ePub ISBN-13: 978-1-59718-310-9 (volumes I)
ePub ISBN-13: 978-1-59718-311-6 (volumes II)
Mobi ISBN-10: 1-59718-312-1 (volumes I and II)
Mobi ISBN-10: 1-59718-313-X (volumes I)
Mobi ISBN-10: 1-59718-314-8 (volumes II)
Mobi ISBN-13: 978-1-59718-312-3 (volumes I and II)
Mobi ISBN-13: 978-1-59718-313-0 (volumes I)
Mobi ISBN-13: 978-1-59718-314-7 (volumes II)

Library of Congress Control Number: 2021944297

Multilevel and Longitudinal Modeling Using Stata

Volume II: Categorical Responses, Counts, and Survival

Fourth Edition

SOPHIA RABE-HESKETH
University of California–Berkeley

ANDERS SKRONDAL
Norwegian Institute of Public Health
University of Oslo
University of California–Berkeley

A Stata Press Publication
StataCorp LLC
College Station, Texas

Published by Stata Press, 4905 Lakeway Drive, College Station, Texas 77845
Typeset in LaTeX 2ε
Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Print ISBN-10: 1-59718-136-6 (volumes I and II)
Print ISBN-10: 1-59718-137-4 (volume I)
Print ISBN-10: 1-59718-138-2 (volume II)
Print ISBN-13: 978-1-59718-136-5 (volumes I and II)
Print ISBN-13: 978-1-59718-137-2 (volume I)
Print ISBN-13: 978-1-59718-138-9 (volume II)
ePub ISBN-10: 1-59718-309-1 (volumes I and II)
ePub ISBN-10: 1-59718-310-5 (volumes I)
ePub ISBN-10: 1-59718-311-3 (volumes II)
ePub ISBN-13: 978-1-59718-309-3 (volumes I and II)
ePub ISBN-13: 978-1-59718-310-9 (volumes I)
ePub ISBN-13: 978-1-59718-311-6 (volumes II)
Mobi ISBN-10: 1-59718-312-1 (volumes I and II)
Mobi ISBN-10: 1-59718-313-X (volumes I)
Mobi ISBN-10: 1-59718-314-8 (volumes II)
Mobi ISBN-13: 978-1-59718-312-3 (volumes I and II)
Mobi ISBN-13: 978-1-59718-313-0 (volumes I)
Mobi ISBN-13: 978-1-59718-314-7 (volumes II)

Library of Congress Control Number: 2021944297

To my children, Astrid and Inge
Anders Skrondal

To Simon
Sophia Rabe-Hesketh

Contents

Contents

(Pages omitted)

Preface

This book is about applied multilevel and longitudinal modeling. Other terms for multilevel models include hierarchical models, random-effects or random-coefficient models, mixed-effects models, or simply mixed models. Longitudinal data are also referred to as panel data, repeated measures, or cross-sectional time series. A popular type of multilevel model for longitudinal data is the growth-curve model.

The common theme of this book is regression modeling when data are clustered in some way. In cross-sectional settings, students may be nested in schools, people in neighborhoods, employees in firms, or twins in twin-pairs. Longitudinal data are by definition clustered because multiple observations over time are nested within units, typically subjects.

Such clustered designs often provide rich information on processes operating at different levels, for instance, people's characteristics interacting with institutional characteristics. Importantly, the standard assumption of independent observations is likely to be violated because of dependence among observations within the same cluster. The multilevel and longitudinal methods discussed in this book extend conventional regression to handle such dependence and exploit the richness of the data.

Volume 1 is on multilevel and longitudinal modeling of continuous responses using linear models. The volume consists of four parts: I. Preliminaries (a review of linear regression modeling, preparing the reader for the rest of the book), II. Two-level models, III. Models for longitudinal and panel data, and IV. Models with nested and crossed random effects. For readers who are new to multilevel and longitudinal modeling, the chapters in part II should be read sequentially and can form the basis of an introductory course on this topic. A one-semester course on multilevel and longitudinal modeling can be based on most of the chapters in volume 1 plus chapter 10 on binary or dichotomous responses from volume 2. For this purpose, we have made chapter 10 freely downloadable from https://www.stata-press.com/books/mlmus4_ch10.pdf.

Volume 2 is on multilevel and longitudinal modeling of categorical responses, counts, and survival data. This volume also consists of four parts: I. Categorical responses (binary or dichotomous responses, ordinal responses, and nominal responses or discrete choice), II. Counts, III. Survival (in both discrete and continuous time), and IV. Models with nested and crossed random effects. Each chapter starts by introducing models for nonclustered data (for example, logistic and Poisson regression) and then extends the models for clustered data by introducing random effects, leading to generalized linear mixed models. Subsequently, alternatives such as generalized estimating equations (GEE) and fixed-effects approaches are discussed. Chapter 10 on binary or dichotomous responses is a core chapter of this volume and should be read before embarking on the other chapters. It is also a good idea to read chapter 14 on discrete-time survival before reading chapter 15 on continuous-time survival.

Our emphasis is on explaining the models and their assumptions, applying the methods to real data, and interpreting results. Many of the issues are conceptually demanding but do not require that you understand complex mathematics. Therefore, wherever possible, we introduce ideas through examples and graphical illustrations, keeping the technical descriptions as simple as possible. Some sections that go beyond an introductory course on multilevel and longitudinal modeling are tagged with the ❖ symbol. Derivations that can be skipped by the reader are given in displays. For an advanced treatment, placing multilevel modeling within a general latent-variable framework, we refer the reader to Skrondal and Rabe-Hesketh (2004), which uses the same notation as this book.

This book shows how all the analyses described can be performed using Stata. There are many advantages of using a general-purpose statistical package such as Stata. First, for those already familiar with Stata, it is convenient not having to learn a new standalone package. Second, conducting multilevel analysis within a powerful package has the advantage that it allows complex data manipulation to be performed, alternative estimation methods to be used, and publication-quality graphics to be produced, all without having to switch packages. Finally, Stata is a natural choice for multilevel and longitudinal modeling because it has gradually become perhaps the most powerful general-purpose statistics package for such models.

Each chapter is based on one or more research problems and real datasets. After describing the models, we walk through the analysis using Stata, pausing to address statistical issues that need further explanation. Do-files for each chapter can be downloaded from https://www.stata-press.com/data/mlmus4.html. Some readers may find it useful to perform the analyses while reading the book.

Stata can be used either via a graphical user interface (GUI) or through commands. We recommend using commands interactively—or preferably in do-files—for serious analysis in Stata. For this reason, and because the GUI is fairly self-explanatory, we use commands exclusively in this book. However, the GUI can be useful for learning the Stata syntax. Generally, we use the `typewriter font` to refer to Stata commands, syntax, and variables. A "dot" prompt followed by a command indicates that you can type verbatim what is displayed after the dot (in context) to replicate the results in the book. Some readers may find it useful to intersperse reading with running these commands. We encourage readers to write do-files for solving the data analysis exercises because this is standard practice for professional data analysis.

The commands used for data manipulation and graphics are explained to some extent, but the purpose of this book is not to teach Stata from scratch. For a basic introduction to Stata, we refer the reader to Acock (2018). Other books and further resources for learning Stata are listed at the Stata website.

If you are new to Stata, we recommend running all the commands given in chapter 1 of volume 1. A list of commands that are particularly useful for manipulating, describing, and plotting multilevel and longitudinal data is given in the appendix of volume 1. Examples using these and other commands can easily be found by referring to the "commands" entry in the subject index.

We have included applications from a wide range of disciplines, including medicine, economics, education, sociology, and psychology. The interdisciplinary nature of this book is also reflected in the choice of models and topics covered. If a chapter is primarily based on an application from one discipline, we try to balance this by including exercises with real data from other disciplines. The two volumes contain over 140 exercises based on over 100 different real datasets. Exercises for which solutions are available to readers are marked with [Solutions], and the solutions can be downloaded from https://www.stata-press.com/books/mlmus4-answers.html. Instructors can obtain solutions to all exercises from Stata Press.

All datasets used in this book are freely available for download; for details, see https://www.stata-press.com/data/mlmus4.html. These datasets can be downloaded into a local directory on your computer. Alternatively, individual datasets can be loaded directly into net-aware Stata by specifying the complete URL. For example,

```
. use https://www.stata-press.com/data/mlmus4/pefr
```

If you have stored the datasets in the working directory, omit the path and just type

```
. use pefr
```

We will generally describe all Stata commands that can be used to fit a given model, discussing their advantages and disadvantages. An exception to this rule is that we do not discuss our own `gllamm` command in volume 1 (see the `gllamm` companion, downloadable from http://www.gllamm.org, for how to fit the models of volume 1 in `gllamm`). In volume 1, we extensively use the Stata commands `xtreg` and `mixed`, and we introduce several more specialized commands for longitudinal modeling, such as `xthtaylor`, `xtivreg`, and `xtabond`. The `sem` command for structural equation modeling is used for growth-curve modeling.

In volume 2, we use Stata's `xt` and `me` commands for different response types. For example, we use `xtlogit` and `melogit` for binary responses, `meologit` for ordinal responses, `xtpoisson` and `mepoisson` for counts, and `mestreg` for multilevel continuous-time survival modeling with shared frailties. In chapter 12 on nominal responses, we use Stata's new `cm` (for "choice model") suite of commands, such as `cmxtmixlogit`. `gllamm` is also used throughout volume 2. We also discuss commands for marginal models and fixed-effects models, such as `xtgee` and `clogit`. The online reference manuals available through the `help` command within Stata provide detailed information on all the official Stata commands for multilevel and longitudinal modeling.

The `nolog` option has been used to suppress the iteration logs showing the progress of the log likelihood. This option is not shown in the command line because we do not recommend it to users; we are using it only to save space.

We assume that readers have a good knowledge of linear regression modeling, in particular, the use and interpretation of dummy variables and interactions. However, the first chapter in volume 1 reviews linear regression and can serve as a refresher.

Errata for different editions and printings of the book can be downloaded from https://www.stata-press.com/books/errata/mlmus4.html, and answers to exercises can be downloaded from https://www.stata-press.com/books/mlmus4-answers.html.

In this fourth edition, we have thoroughly revised all chapters and updated the Stata syntax for release 17. Major additions in volume 1 include the Kenward–Roger degree-of-freedom correction for improved inference with a small number of clusters, difference-in-difference estimation for quasi experiments, and instrumental-variables estimation to handle level-1 endogeneity. In volume 2, we now introduce Bayesian estimation for crossed-effects models and extensively use several new commands (since the third edition), including `meologit`, `cmxtmixlogit`, `mestreg`, and `menbreg`.

Berkeley and Oslo
August 2021

Sophia Rabe-Hesketh
Anders Skrondal

(Pages omitted)

4 Random-coefficient models

4.1 Introduction

In the previous chapter, we considered linear random-intercept models where the overall level of the response was allowed to vary between clusters after controlling for covariates. In this chapter, we include random coefficients or random slopes in addition to random intercepts, thus also allowing the effects of covariates to vary between clusters. Such models involving both random intercepts and random slopes are often called *random-coefficient models*. In longitudinal settings, where the level-1 units are occasions and the clusters are typically subjects, models with a random-coefficient of time are also referred to as growth-curve models. Such models are the topic of chapter 7.

4.2 How effective are different schools?

Here we analyze a dataset on inner-London schools that accompanies the MLwiN software (Rasbash et al. 2019) and is part of the data analyzed by Goldstein et al. (1993).

At age 16, students took their Graduate Certificate of Secondary Education (GCSE) exams in a number of subjects. A score was derived from the individual exam results. Such scores often form the basis for school comparisons, for instance, to allow parents to choose the best school for their child. However, schools can differ considerably in their intake achievement levels. It may be argued that what should be compared is the "value added"; that is, the difference in mean GCSE score between schools after controlling for the students' achievement before entering the school. One such measure of prior achievement is the London Reading Test (LRT) taken by these students at age 11.

The dataset `gcse.dta` has the following variables:

- `school`: school identifier
- `student`: student identifier
- `gcse`: Graduate Certificate of Secondary Education (GCSE) score (z score, multiplied by 10)
- `lrt`: London Reading Test (LRT) score (z score, multiplied by 10)
- `girl`: dummy variable for student being a girl (1: girl; 0: boy)
- `schgend`: type of school (1: mixed gender; 2: boys only; 3: girls only)

One purpose of the analysis is to investigate the relationship between GCSE and LRT and how this relationship varies between schools. The model can then be used to address the question of which schools appear to be most effective, taking prior achievement into account.

We read in the data by using

```
. use https://www.stata-press.com/data/mlmus4/gcse
```

4.3 Separate linear regressions for each school

Before developing a model for all 65 schools combined, we consider a separate model for each school. For school j, an obvious model for the relationship between GCSE and LRT is a simple regression model,

$$y_{ij} \;=\; \beta_{1j} + \beta_{2j} x_{ij} + \epsilon_{ij}$$

where y_{ij} is the GCSE score for the ith student in school j, x_{ij} is the corresponding LRT score, β_{1j} is the school-specific intercept, β_{2j} is the school-specific slope, and ϵ_{ij} is a residual error term with school-specific variance θ_j.

For school 1, OLS estimates of the intercept $\widehat{\beta}_{11}$ and the slope $\widehat{\beta}_{21}$ can be obtained using `regress`,

```
. regress gcse lrt if school==1
      Source |       SS           df       MS      Number of obs   =        73
-------------+----------------------------------   F(1, 71)        =     59.44
       Model |  4084.89189         1  4084.89189   Prob > F        =    0.0000
    Residual |  4879.35759        71  68.7233463   R-squared       =    0.4557
-------------+----------------------------------   Adj R-squared   =    0.4480
       Total |  8964.24948        72  124.503465   Root MSE        =      8.29

------------------------------------------------------------------------------
        gcse | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         lrt |   .7093406   .0920061     7.71   0.000     .5258856    .8927955
       _cons |   3.833302   .9822377     3.90   0.000     1.874776    5.791828
------------------------------------------------------------------------------
```

where we have selected school 1 by specifying the condition `if school==1`.

To assess whether this is a reasonable model for school 1, we can obtain the predicted (ordinary least squares) regression line for this school (with $j = 1$),

$$\widehat{y}_{i1} = \widehat{\beta}_{11} + \widehat{\beta}_{21} x_{i1}$$

by using the `predict` command with the `xb` option:

```
. predict p_gcse, xb
```

We superimpose this line on the scatterplot of the data for the school, as shown in figure 4.1.

```
. twoway (scatter gcse lrt) (line p_gcse lrt, sort) if school==1,
> xtitle(LRT) ytitle(GCSE)
```

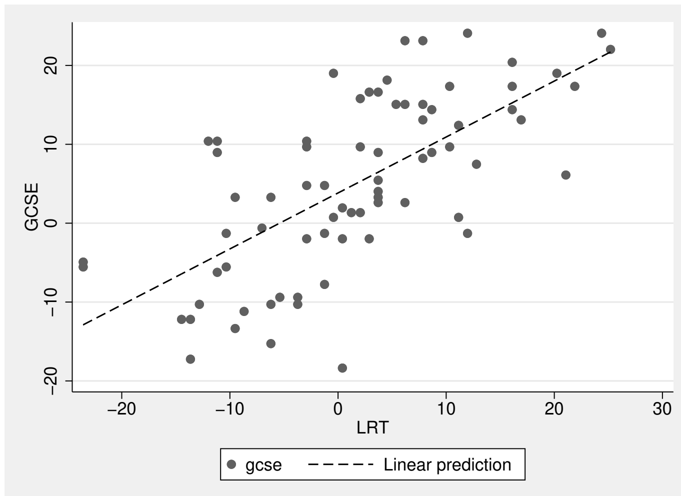

Figure 4.1: Scatterplot of `gcse` versus `lrt` for school 1 with ordinary least-squares regression line

We can also produce a *trellis graph* containing such plots for all 65 schools by using

```
. twoway (scatter gcse lrt) (lfit gcse lrt, sort lpatt(solid)),
> by(school, compact legend(off) cols(5))
> xtitle(LRT) ytitle(GCSE) ysize(3) xsize(2)
```

where `lfit` plots regression lines estimated by OLS. The resulting graph, shown in figure 4.2, suggests that the model assumptions are reasonably met. The schools appear to vary in both their intercepts and slopes.

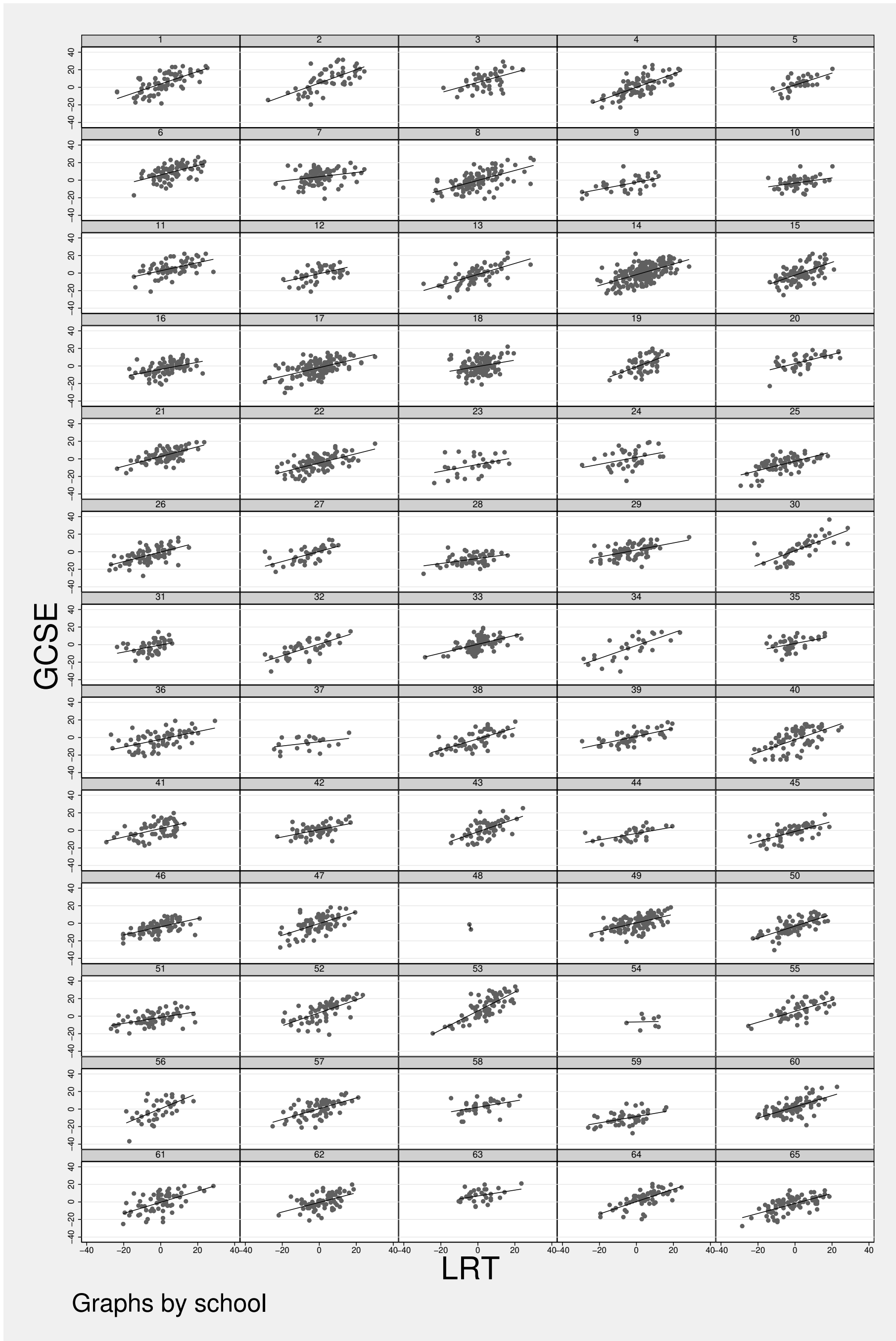

Figure 4.2: Trellis of scatterplots of `gcse` versus `lrt` with fitted regression lines for all 65 schools

We will now fit a simple linear regression model for each school, which is easily done using Stata's prefix command statsby. Then we will examine the variability in the estimated intercepts and slopes.

First, calculate the number of students per school by using egen with the count() function to preclude fitting lines to schools with fewer than five students below:

```
. egen num = count(gcse), by(school)
```

Then, use statsby to create a new dataset, ols.dta, in the working directory with the variables inter and slope containing OLS estimates of the intercepts (_b[_cons]) and slopes (_b[lrt]) from the command regress gcse lrt if num>4 applied to each school:

```
. statsby inter=_b[_cons] slope=_b[lrt], by(school) saving(ols):
> regress gcse lrt if num>4
(running regress on estimation sample)

      Command: regress gcse lrt if num>4
        inter: _b[_cons]
        slope: _b[lrt]
           By: school

Statsby groups
────┼─── 1 ───┼─── 2 ───┼─── 3 ───┼─── 4 ───┼─── 5
..................................................    50
.............
```

The new dataset also contains the variable school and is sorted by school, making it easy to merge it into the original dataset (the "master data") after sorting the latter by school:

```
. sort school

. merge m:1 school using ols

    Result                           Number of obs
    ─────────────────────────────────────────────
    Not matched                                 2
        from master                             2  (_merge==1)
        from using                              0  (_merge==2)

    Matched                                 4,057  (_merge==3)
    ─────────────────────────────────────────────

. drop _merge
```

Here we have specified m:1 in the merge command, which stands for "many-to-one merging" (observations for several students per school in the master data, but only one observation per school in the "using data"). We see that two of the students in the master data did not have matches in the using data (because their school, school 48, had fewer than 5 students in the data, so we did not compute OLS estimates for that school). We have deleted the variable _merge produced by the merge command to avoid error messages when we run the merge command in the future.

A scatterplot of the OLS estimates of the intercept and slope is produced using the following command and given in figure 4.3:

```
. twoway scatter slope inter, xtitle(Intercept) ytitle(Slope)
```

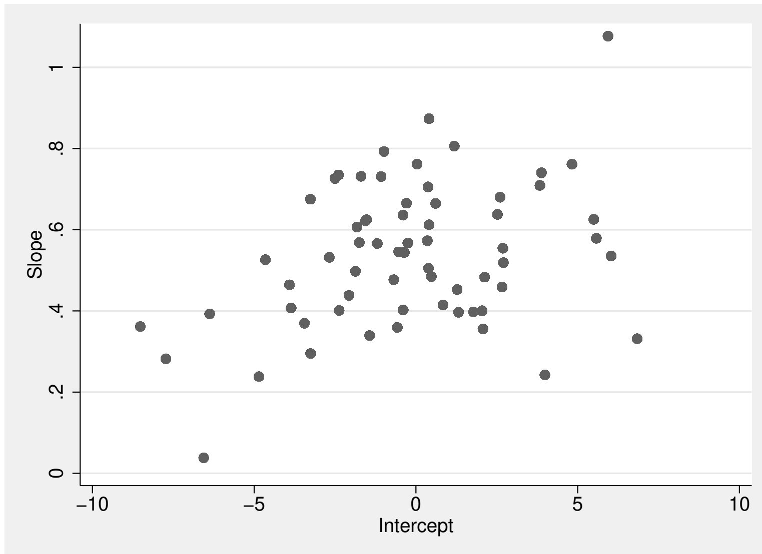

Figure 4.3: Scatterplot of estimated intercepts and slopes for all schools with at least five students

We see that there is considerable variability between the estimated intercepts and slopes of different schools. To investigate this further, we first create a dummy variable to pick out one observation per school,

```
. egen pickone = tag(school)
```

and then produce summary statistics for the schools by using the `summarize` command:

```
. summarize inter slope if pickone==1
    Variable         Obs        Mean    Std. dev.       Min        Max

       inter          64   -.1805974    3.291357  -8.519253   6.838716
       slope          64    .5390514    .1766135   .0380965   1.076979
```

To allow comparison with the parameter estimates obtained from the random-coefficient model considered later on, we also obtain the covariance matrix of the estimated intercepts and slopes:

```
. correlate inter slope if pickone==1, covariance
(obs=64)

                inter    slope

       inter   10.833
       slope  .208622  .031192
```

The diagonal elements, 10.83 and 0.03, are the sample variances of the intercepts and slopes, respectively. The off-diagonal element, 0.21, is the sample covariance between the intercepts and slopes, equal to the correlation times the product of the intercept and slope standard deviations.

We can also obtain a *spaghetti plot* of the predicted school-specific regression lines for all schools. We first calculate the fitted values $\widehat{y}_{ij} = \widehat{\beta}_{1j} + \widehat{\beta}_{2j}x_{ij}$,

```
. generate pred = inter + slope*lrt
(2 missing values generated)
```

and sort the data so that `lrt` increases within a given school and then jumps to its lowest value for the next school in the dataset:

```
. sort school lrt
```

We then produce the plot by typing

```
. twoway (line pred lrt, connect(ascending)), xtitle(LRT)
> ytitle(Fitted regression lines)
```

The `connect(ascending)` option is used to connect points only as long as `lrt` is increasing and ensures that only data for the same school are connected. The resulting graph is shown in figure 4.4.

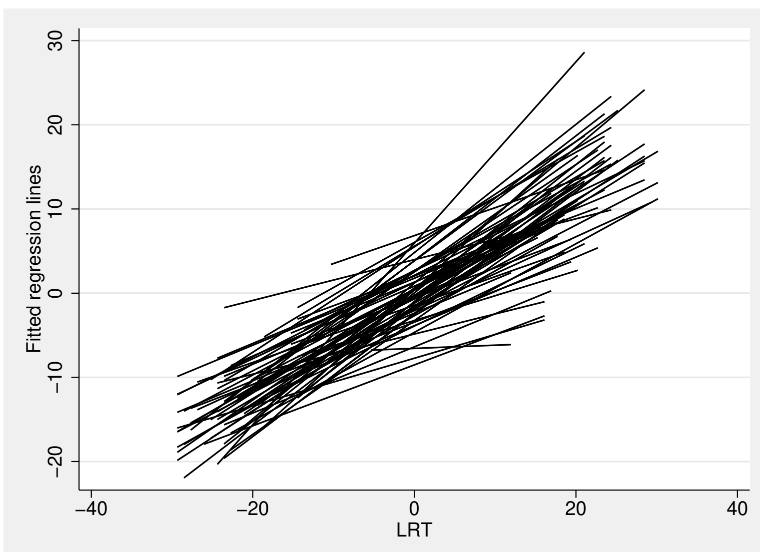

Figure 4.4: Spaghetti plot of ordinary least-squares regression lines for all schools with at least five students

4.4 Specification and interpretation of a random-coefficient model

4.4.1 Specification of a random-coefficient model

How can we develop a joint model for the relationships between `gcse` and `lrt` in all schools that allows intercepts and slopes to differ between schools?

One way would be to use dummy variables for all schools (omitting the overall intercept) and interactions between these dummy variables and `lrt` (omitting the overall slope of `lrt`). The school-specific intercepts are then the coefficients of the dummy variables and the school-specific slopes are the interaction coefficients. The only difference between the resulting model and separate regressions is that a common residual error variance $\theta_j = \theta$ is assumed. However, this model has 130 regression coefficients! Furthermore, if the schools are viewed as a (random) sample of schools from a population of schools, we are not interested in the individual coefficients characterizing each school's regression line. Rather, we would like to estimate the mean intercept and slope as well as the (co)variability of the intercepts and slopes in the population of schools.

A parsimonious model for the relationships between `gcse` and `lrt` can be obtained by specifying a school-specific random intercept ζ_{1j} and a school-specific random slope ζ_{2j} for `lrt` (x_{ij}):

$$\begin{aligned} y_{ij} &= \beta_1 + \beta_2 x_{ij} + \zeta_{1j} + \zeta_{2j} x_{ij} + \epsilon_{ij} \\ &= (\beta_1 + \zeta_{1j}) + (\beta_2 + \zeta_{2j}) x_{ij} + \epsilon_{ij} \end{aligned} \tag{4.1}$$

Here ζ_{1j} represents the deviation of school j's intercept from the mean intercept β_1, and ζ_{2j} represents the deviation of school j's slope from the mean slope β_2.

Given all covariates $\mathbf{X}_j$ in cluster j, it is assumed that the random effects ζ_{1j} and ζ_{2j} have zero expectations:

$$E(\zeta_{1j}|\mathbf{X}_j) = 0$$

$$E(\zeta_{2j}|\mathbf{X}_j) = 0$$

It is also assumed that the level-1 residual ϵ_{ij} has zero expectation, given the covariates and the random effects:

$$E(\epsilon_{ij}|\mathbf{X}_j, \zeta_{1j}, \zeta_{2j}) = 0$$

It follows from these mean-independence assumptions that the random terms ζ_{1j}, ζ_{2j}, and ϵ_{ij} are all uncorrelated with the covariate x_{ij} and with $\overline{x}_{\cdot j}$ and that ϵ_{ij} is uncorrelated with both ζ_{1j} and ζ_{2j}. Both the intercepts ζ_{1j} and slopes ζ_{2j} are assumed to be uncorrelated across schools, and the level-1 residuals ϵ_{ij} are assumed to be uncorrelated across schools and students.

An illustration of this random-coefficient model with one covariate x_{ij} for one cluster j is shown in the bottom panel of figure 4.5. A random-intercept model is shown for comparison in the top panel.

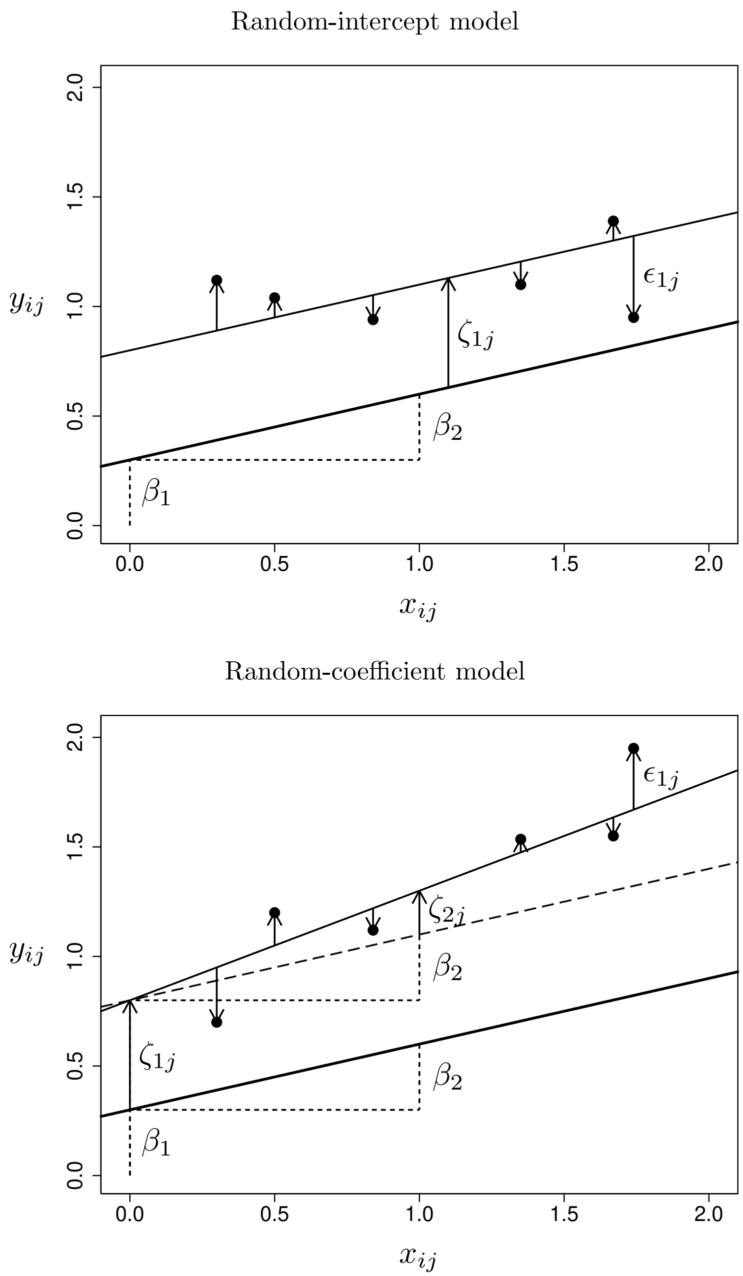

Figure 4.5: Illustration of random-intercept and random-coefficient models

In each panel, the lower bold and solid line represents the population-averaged or marginal regression line

$$E(y_{ij}|x_{ij}) = \beta_1 + \beta_2 x_{ij}$$

across all clusters. The higher and thinner solid line represents the cluster-specific regression line for cluster j. The arrows from the cluster-specific regression lines to the responses y_{ij} are the within-cluster residual error terms ϵ_{ij} (with variance θ).

For the random-intercept model, the cluster-specific line is

$$E(y_{ij}|x_{ij}, \zeta_{1j}) = (\beta_1 + \zeta_{1j}) + \beta_2 x_{ij}$$

which is parallel to the population-averaged line with vertical displacement given by the random intercept ζ_{1j}. In contrast, in the random-coefficient model, the cluster-specific or conditional regression line

$$E(y_{ij}|x_{ij}, \zeta_{1j}, \zeta_{2j}) = (\beta_1 + \zeta_{1j}) + (\beta_2 + \zeta_{2j})x_{ij}$$

is not parallel to the population-averaged line but has a greater slope because the random slope ζ_{2j} is positive in the illustration. Here the dashed line is parallel to the population-averaged regression line and has the same intercept as cluster j. The vertical deviation between this dashed line and the line for cluster j is $\zeta_{2j}x_{ij}$, as shown in the diagram for $x_{ij}=1$. The bottom panel illustrates that the total intercept for cluster j is $\beta_1 + \zeta_{1j}$ and the total slope is $\beta_2 + \zeta_{2j}$. It is clear that $\zeta_{2j}x_{ij}$ represents an *interaction* between the clusters, treated as random, and the covariate x_{ij}.

Given $\mathbf{X}_j$, the random intercept and random slope have a bivariate distribution assumed to have 0 means and covariance matrix $\boldsymbol{\Psi}$:

$$\boldsymbol{\Psi} = \begin{bmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{bmatrix} \equiv \begin{bmatrix} \mathrm{Var}(\zeta_{1j}|\mathbf{X}_j) & \mathrm{Cov}(\zeta_{1j}, \zeta_{2j}|\mathbf{X}_j) \\ \mathrm{Cov}(\zeta_{2j}, \zeta_{1j}|\mathbf{X}_j) & \mathrm{Var}(\zeta_{2j}|\mathbf{X}_j) \end{bmatrix}, \qquad \psi_{21} = \psi_{12}$$

Hence, given the covariates, the variance of the random intercept is ψ_{11}, the variance of the random slope is ψ_{22}, and the covariance between the random intercept and the random slope is ψ_{21}. The correlation between the random intercept and random slope given the covariates becomes

$$\rho_{21} \equiv \mathrm{Cor}(\zeta_{1j}, \zeta_{2j}|\mathbf{X}_j) \;=\; \frac{\psi_{21}}{\sqrt{\psi_{11}\psi_{22}}}$$

For maximum likelihood (ML) and restricted maximum likelihood (REML) estimation a normal distribution is specified for the level-1 error ϵ_{ij} and a bivariate normal distribution for the random intercept and random slope, given $\mathbf{X}_j$. An example of a bivariate normal distribution with $\psi_{11} = \psi_{22} = 4$ and $\psi_{21} = \psi_{12} = 1$ is shown as a perspective plot in figure 4.6. Specifying a bivariate normal distribution implies that the (marginal) univariate distributions of the intercept and slope are also normal.

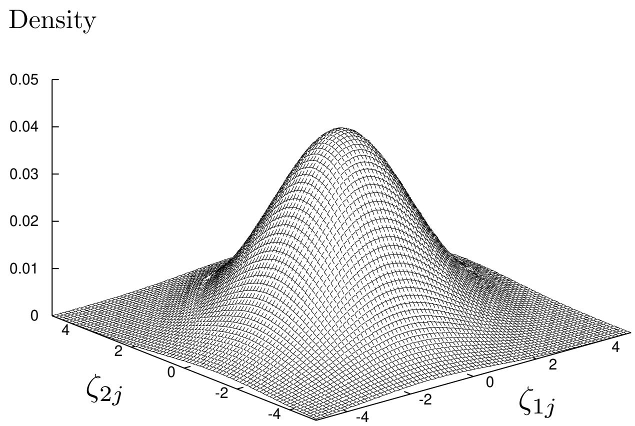

Figure 4.6: Perspective plot of bivariate normal distribution

4.4.2 Interpretation of the random-effects variances and covariances

Interpreting the covariance matrix $\mathbf{\Psi}$ of the random effects (given the covariates $\mathbf{X}_j$) is not completely straightforward.

First, the random-slope variance ψ_{22} and the covariance between random slope and intercept ψ_{21} depend not just on the scale of the response variable but also on the scale of the covariate, here `lrt`. Let the units of the response and covariate be denoted as u_y and u_x, respectively. For instance, in an application in chapter 7 that considers children's increase in weight over time, u_y is kilograms and u_x is years. The units of ψ_{11} are u_y^2, the units of ψ_{21} are u_y^2/u_x, and the units of ψ_{22} are u_y^2/u_x^2. It therefore does not make sense to compare the magnitude of random-intercept and random-slope variances.

Another issue is that the total residual variance is no longer constant as in random-intercept models. The total residual is now

$$\xi_{ij} \equiv \zeta_{1j} + \zeta_{2j}x_{ij} + \epsilon_{ij}$$

and the conditional variance of the responses given the covariate, or the conditional variance of the total residual, is

$$\text{Var}(y_{ij}|\mathbf{X}_j) = \text{Var}(\xi_{ij}|\mathbf{X}_j) = \psi_{11} + 2\psi_{21}x_{ij} + \psi_{22}x_{ij}^2 + \theta \tag{4.2}$$

This variance is a (quadratic) function of the covariate x_{ij}, and the total residual is therefore *heteroskedastic*. The conditional covariance for two students i and i' with covariate values x_{ij} and $x_{i'j}$ in the same school j is

$$\begin{aligned} \text{Cov}(y_{ij}, y_{i'j}|\mathbf{X}_j) &= \text{Cov}(\xi_{ij}, \xi_{i'j}|\mathbf{X}_j) \\ &= \psi_{11} + \psi_{21}x_{ij} + \psi_{21}x_{i'j} + \psi_{22}x_{ij}x_{i'j} \end{aligned} \tag{4.3}$$

and the conditional intraclass correlation becomes

$$\text{Cor}(y_{ij}, y_{i'j}|\mathbf{X}_j) = \frac{\text{Cov}(\xi_{ij}, \xi_{i'j}|\mathbf{X}_j)}{\sqrt{\text{Var}(\xi_{ij}|\mathbf{X}_j)\text{Var}(\xi_{i'j}|\mathbf{X}_j)}}$$

where we can plug in the covariance from (4.3) and the variances from (4.2). When $x_{ij}=x_{i'j}=0$, the expression for the intraclass correlation is the same as for the random-intercept model and represents the correlation of the total residuals (from the overall mean regression line) for two students in the same school who both have `lrt` scores equal to 0 (the mean in this case). However, for pairs of students i and i' in the same school j with other values of `lrt`, the intraclass correlation is a complicated function of `lrt` (x_{ij} and $x_{i'j}$).

Due to the heteroskedastic total residual variance, it is not straightforward to define coefficients of determination—such as R^2, R_2^2, and R_1^2, discussed in section 3.5—for random-coefficient models. Snijders and Bosker (2012, 114) suggest removing the random coefficient(s) for the purpose of calculating the coefficient of determination because this will usually yield values that are close to correct (see their section 7.2.2 for how to obtain the correct version).

Finally, interpreting the parameters ψ_{11} and ψ_{21} can be difficult because their values depend on the translation of the covariate or, in other words, on how much we add or subtract from the covariate. Adding a constant to `lrt` and refitting the model would result in different estimates of ψ_{11} and ψ_{21} (see also exercise 4.9). This is because the intercept variance is the variability in the vertical positions of school-specific regression lines where `lrt`=0 (the position where `lrt`=0 changes when `lrt` is translated) and the covariance or correlation is the tendency for regression lines that are higher up where `lrt`=0 to have higher slopes. This lack of invariance of ψ_{11} and ψ_{21} to translation of the covariate x_{ij} is illustrated in figure 4.7. Here identical cluster-specific regression lines are shown in the two panels, but the covariate $x'_{ij} = x_{ij} - 3.5$ in the lower panel is translated relative to the covariate x_{ij} in the upper panel. The intercepts are the intersections of the regression lines with the vertical lines at 0. Clearly these intercepts vary more in the upper panel than the lower panel, whereas the correlation between intercepts and slopes is negative in the upper panel and positive in the lower panel.

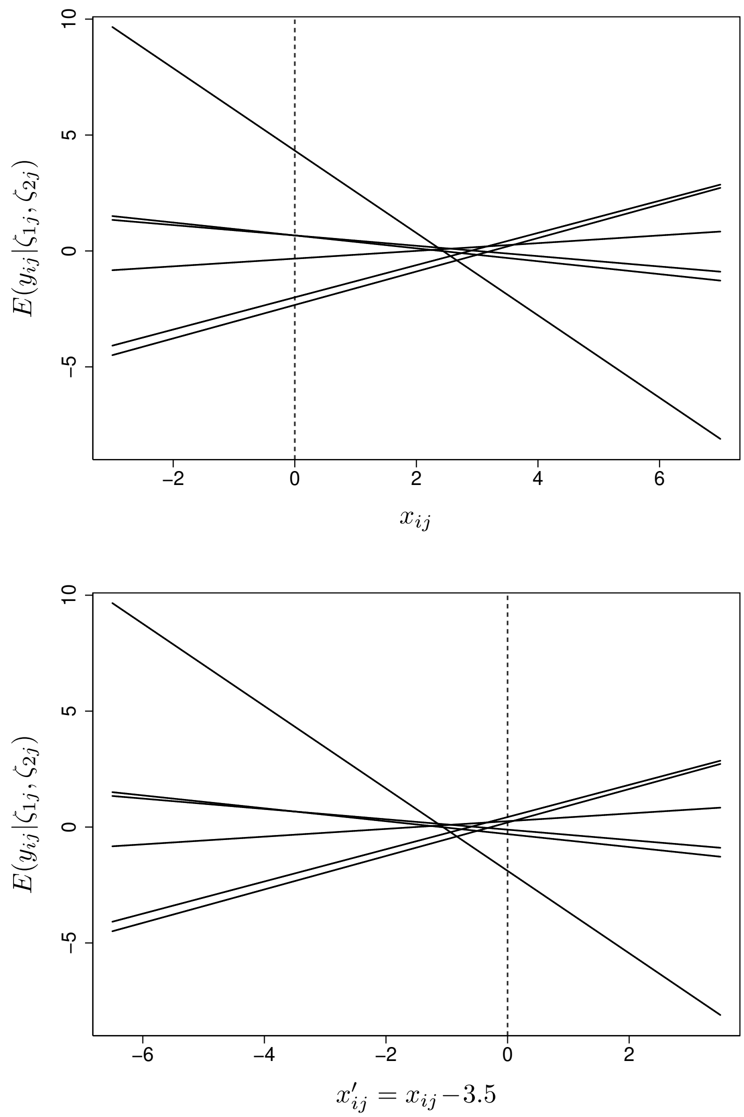

Figure 4.7: Cluster-specific regression lines for random-coefficient model, illustrating lack of invariance under translation of covariate (*Source:* Skrondal and Rabe-Hesketh 2004)

To make ψ_{11} and ψ_{21} interpretable, it makes sense to translate x_{ij} so that the value $x_{ij} = 0$ is a useful reference point in some way. Typical choices are either mean centering (as for `lrt`) or, if x_{ij} is time, as in growth-curve models, defining 0 to be the initial time in some sense. Because the magnitude and interpretation of ψ_{21} depend on the location (or translation) of x_{ij}, which is often arbitrary, it generally does not make sense to set ψ_{21} to 0 by specifying uncorrelated intercepts and slopes.

A useful way of interpreting the magnitudes of the estimated variances $\widehat{\psi}_{11}$ and $\widehat{\psi}_{22}$ is by constructing intervals that contain the intercepts and slopes of 95% of clusters in the population (treating estimates as known parameters). Assuming that the intercepts and slopes are normally distributed with means $\widehat{\beta}_1$ and $\widehat{\beta}_2$ and variances $\widehat{\psi}_{11}$ and $\widehat{\psi}_{22}$, these intervals are $\widehat{\beta}_1 \pm 1.96\sqrt{\widehat{\psi}_{11}}$ and $\widehat{\beta}_2 \pm 1.96\sqrt{\widehat{\psi}_{22}}$. To aid interpretation of the random part of the model, it is also useful to produce plots of predicted school-specific regression lines, as discussed in section 4.8.3.

4.5 Estimation using mixed

The `mixed` command can be used to fit linear random-coefficient models by ML or REML. (`xtreg` can only fit two-level random-intercept models.)

4.5.1 Random-intercept model

We first consider a random-intercept model discussed in the previous chapter:

$$y_{ij} = (\beta_1 + \zeta_{1j}) + \beta_2 x_{ij} + \epsilon_{ij}$$

This model is a special case of the random-coefficient model in (4.1) with $\zeta_{2j} = 0$ or, equivalently, with zero random-slope variance and zero random-intercept and random-slope covariance, $\psi_{22} = \psi_{21} = 0$.

ML estimates for the random-intercept model can be obtained using `mixed` with the `mle` option (the default), and we also use the `vce(robust)` option for robust standard errors:

```
. mixed gcse lrt || school:, mle stddeviations vce(robust)
Mixed-effects regression                        Number of obs     =      4,059
Group variable: school                          Number of groups  =         65
                                                Obs per group:
                                                              min =          2
                                                              avg =       62.4
                                                              max =        198
                                                Wald chi2(1)      =     852.73
Log pseudolikelihood = -14024.799               Prob > chi2       =     0.0000
                                (Std. err. adjusted for 65 clusters in school)

                             Robust
        gcse   Coefficient  std. err.      z    P>|z|     [95% conf. interval]

         lrt     .5633697   .0192925    29.20   0.000     .5255572    .6011823
       _cons     .0238706   .4050143     0.06   0.953    -.7699428    .8176841


                                            Robust
  Random-effects parameters    Estimate   std. err.     [95% conf. interval]

school: Identity
                  sd(_cons)    3.035269   .3154741      2.475863     3.72107

               sd(Residual)    7.521481   .1306016      7.269813    7.781861
```

To allow later comparison with random-coefficient models via likelihood-ratio tests, we store these estimates by using

```
. estimates store ri
```

The random-intercept model assumes that the school-specific regression lines are parallel. The common coefficient or slope β_2 of `lrt`, shared by all schools, is estimated as 0.56 and the mean intercept as 0.02. Schools vary in their intercepts with an estimated standard deviation of 3.04. Within the schools, the estimated residual standard deviation around the school-specific regression lines is 7.52. The within-school correlation, after controlling for `lrt`, is therefore estimated as

$$\widehat{\rho} \;=\; \frac{\widehat{\psi}_{11}}{\widehat{\psi}_{11}+\widehat{\theta}} \;=\; \frac{3.035^2}{3.035^2+7.521^2} = 0.14$$

We could obtain this within-school correlation by typing `estat icc`.

The ML estimates for the random-intercept model are also given under “Random intercept” in table 4.1.

Table 4.1: Maximum likelihood estimates for inner-London-schools data with robust standard errors

	Random intercept		Random coefficient		Rand. coefficient & level-2 covariates		
Parameter	Est	(SE)	Est	(SE)	Est	(SE)	γ_{xx}
Fixed part							
β_1 [_cons]	0.02	(0.41)	−0.12	(0.40)	−1.00	(0.55)	γ_{11}
β_2 [lrt]	0.56	(0.02)	0.56	(0.02)	0.57	(0.02)	γ_{21}
β_3 [boys]					0.85	(0.96)	γ_{12}
β_4 [girls]					2.43	(0.84)	γ_{13}
β_5 [boys_lrt]					−0.02	(0.05)	γ_{22}
β_6 [girls_lrt]					−0.03	(0.05)	γ_{23}
Random part							
$\sqrt{\psi_{11}}$	3.04		3.01		2.80		
$\sqrt{\psi_{22}}$			0.12		0.12		
ρ_{21}			0.50		0.60		
$\sqrt{\theta}$	7.52		7.44		7.44		
Log likelihood	$-14,024.80$		$-14,004.61$		$-13,998.83$		

4.5.2 Random-coefficient model

We now relax the assumption that the school-specific regression lines are parallel by introducing random school-specific slopes $\beta_2 + \zeta_{2j}$ of `lrt`:

$$y_{ij} = (\beta_1 + \zeta_{1j}) + (\beta_2 + \zeta_{2j})x_{ij} + \epsilon_{ij}$$

To introduce a random slope for `lrt` using `mixed`, we simply add that variable name in the specification of the random part, replacing `school:` with `school: lrt`. We must also specify the `covariance(unstructured)` option because `mixed` will otherwise set the covariance ψ_{21} (and the corresponding correlation) to 0 by default. ML estimates for the random-coefficient model are then obtained using

```
. mixed gcse lrt || school: lrt, covariance(unstructured) mle stddeviations
> vce(robust)
Mixed-effects regression                        Number of obs     =      4,059
Group variable: school                          Number of groups  =         65
                                                Obs per group:
                                                              min =          2
                                                              avg =       62.4
                                                              max =        198
                                                Wald chi2(1)      =     767.80
Log pseudolikelihood = -14004.613               Prob > chi2       =     0.0000
                                (Std. err. adjusted for 65 clusters in school)
```

gcse	Coefficient	Robust std. err.	z	P>|z|	[95% conf. interval]	
lrt	.556729	.0200919	27.71	0.000	.5173496	.5961084
_cons	-.115085	.4009294	-0.29	0.774	-.9008922	.6707222

Random-effects parameters	Estimate	Robust std. err.	[95% conf. interval]	
school: Unstructured				
sd(lrt)	.1205646	.0236268	.0821128	.1770224
sd(_cons)	3.007444	.3134589	2.451765	3.689065
corr(lrt,_cons)	.4975415	.1751841	.0894796	.7625783
sd(Residual)	7.440787	.1251535	7.19949	7.690172

Because the `stddeviations` option was used, the output shows the standard deviations, `sd(lrt)`, of the slope and `sd(_cons)` of the intercept instead of variances. It also shows the correlation between intercepts and slopes, `corr(lrt,_cons)`, instead of the covariance. We can obtain the estimated covariance matrix either by replaying the estimation results without the `stddeviations` option (or with the `variance` option),

```
mixed, variance
```

or by using the postestimation command `estat recovariance`:

```
. estat recovariance
Random-effects covariance matrix for level school
```

	lrt	_cons
lrt	.0145358	
_cons	.1804042	9.04472

The ML estimates for the random-coefficient model were also given under "Random coefficient" in table 4.1. We store the estimates under the name `rc` for later use:

```
. estimates store rc
```

We can also obtain the model-implied residual standard deviations and correlations among the GCSE scores for students in a particular school by using the `estat` wcorrelation command. For schools that have many students in the data, the correlation matrix is too large to display without wrapping, so we choose school 54, which has 8 students in the data, for illustration. First, we sort the data in ascending order of lrt within school and list the values of lrt because they will affect both the standard deviations and correlations, as shown in equations (4.3) and (4.2):

```
. sort school lrt
. list school lrt if school==54, clean noobs
   school       lrt
       54   -5.3806
       54     2.058
       54    2.8845
       54     3.711
       54    9.4967
       54    10.323
       54    11.976
       54    11.976
```

Now, we obtain the estimated residual standard deviations and correlations for school 54:

```
. estat wcorrelation, at(school=54)
Standard deviations and correlations for school = 54:
Standard deviations:
         obs |     1       2       3       4       5       6       7       8
-------------+----------------------------------------------------------------
          sd | 7.930   8.076   8.098   8.121   8.315   8.348   8.415   8.415
Correlations:
         obs |     1       2       3       4       5       6       7       8
-------------+----------------------------------------------------------------
           1 | 1.000
           2 | 0.129   1.000
           3 | 0.130   0.153   1.000
           4 | 0.131   0.155   0.158   1.000
           5 | 0.137   0.170   0.173   0.177   1.000
           6 | 0.138   0.172   0.175   0.179   0.202   1.000
           7 | 0.139   0.176   0.179   0.183   0.208   0.212   1.000
           8 | 0.139   0.176   0.179   0.183   0.208   0.212   0.218   1.000
```

The standard deviations increase with increasing lrt. To interpret the pattern of the correlations, we can look down the columns, which corresponds to holding the lrt for one student constant and looking at the correlations as the lrt of the other student increases. We see that the corresponding correlations increase.

Here we used ML estimation. REML estimation should be used instead when the number of clusters is small ($J - q < 42$, see display 2.1) and this method is requested by specifying the reml option.

4.6 Testing the slope variance

Before interpreting the parameter estimates, we may want to test whether the random slope is needed in addition to the random intercept. Specifically, we test the null hypothesis

$$H_0\colon\ \psi_{22} = 0 \quad \text{against} \quad H_a\colon \psi_{22} > 0$$

Note that H_0 is equivalent to the hypothesis that the random slopes ζ_{2j} are all 0. The null hypothesis also implies that $\psi_{21} = 0$, because a variable that does not vary also does not covary with other variables. Setting $\psi_{22} = 0$ and $\psi_{21} = 0$ gives the random-intercept model.

A naïve likelihood-ratio test can be performed using the `lrtest` command:

```
. lrtest ri rc, force
Likelihood-ratio test
Assumption: ri nested within rc
 LR chi2(2) =  40.37
Prob > chi2 = 0.0000
```

The `force` option was used here because without it, Stata will not perform likelihood-ratio tests when robust standard errors have been specified. This is because inferences based on robust standard errors do not require the likelihood to be correct (that is, to correspond to the data-generating mechanism), which is why Stata calls it a pseudolikelihood in the output. Because likelihood-ratio tests require correct likelihoods, Stata will not perform such tests unless forced to do so. Here we accept that, unlike inferences for the regression coefficients based on robust standard errors, likelihood-ratio tests for variance and covariance parameters will not be robust to misspecification of the residual covariance structure. Remember that point estimators of variance and covariance parameters are inconsistent if the residual covariance structure is misspecified (which is also unlike regression coefficients).

This likelihood-ratio test is naïve because the variance ψ_{22} must be nonnegative so that the null hypothesis is on the boundary of the parameter space. As discussed in section 2.6.2 for random-intercept models, the asymptotic null distribution of the likelihood-ratio statistic L is therefore no longer a simple χ^2 distribution as assumed by the `lrtest` command.

In `mixed`, the default estimation metric (transformation used during estimation) for the covariance matrix of the random effects is the square root or Cholesky decomposition (which is requested by the `matsqrt` option). This parameterization forces the covariance matrix to be *positive semidefinite* (estimates on the boundary of parameter space, for example, 0 variance or perfect correlations, are allowed). It can be shown that the asymptotic null distribution for testing the null hypothesis that the variance of the $r{+}1$th random effect is 0 becomes $0.5\,\chi^2(r) + 0.5\,\chi^2(r+1)$. For our case of testing the random slope variance in a model with a random intercept and a random slope, $r = 1$; it follows that the asymptotic null distribution is $0.5\,\chi^2(1) + 0.5\,\chi^2(2)$. The correct p-value can be obtained as

```
. display 0.5*chi2tail(1,40.37) + 0.5*chi2tail(2,40.37)
9.616e-10
```

We see that the conclusion remains the same as for the naïve approach for this application.

If the `matlog` option is used, the estimation metric for the covariance matrix of the random effects is matrix logarithms, which forces the covariance matrix to be *positive definite* (estimates on the boundary of the parameter space are not allowed). Consequently, convergence is not achieved if the ML estimates are on the boundary of the parameter space. If this leads to reverting to the model under the null hypothesis, giving a likelihood-ratio statistic equal to 0, then the asymptotic null distribution for testing the null hypothesis that the variance of the $r+1$th random effect is 0 becomes $0.5\,\chi^2(0) + 0.5\,\chi^2(r+1)$, where $\chi^2(0)$ has a probability mass of 1 at 0. For testing the random slope variance in a model with a random intercept and a random slope, $r = 1$ and the distribution becomes $0.5\,\chi^2(0)+0.5\,\chi^2(2)$ so that the correct p-value can simply be obtained by dividing the naïve p-value based on the $\chi^2(2)$ by 2.

Keep in mind that the naïve likelihood-ratio test for testing the slope variance is conservative. Hence, if the null hypothesis of a zero slope variance is rejected by the naïve approach, it is also rejected by the correct approach.

Unfortunately, there is no straightforward procedure available for testing several variances simultaneously, unless the random effects are independent (see section 8.8), and simulations (for example, parametric bootstrapping) must be used in this case to obtain the empirical null distribution.

4.7 Interpretation of estimates

The population-mean intercept and slope are estimated as -0.12 and 0.56, respectively. These estimates are similar to those for the random-intercept model (see table 4.1) and are also close to the means of the school-specific intercept and slope estimates given in section 4.3.

The estimated random-intercept standard deviation and level-1 residual standard deviation are somewhat lower than for the random-intercept model. The latter is because of a better fit of the school-specific regression lines for the random-coefficient model, which relaxes the restriction of parallel regression lines. The estimated covariance matrix of the intercepts and slopes is similar to the sample covariance matrix of the ordinary least-squares estimates reported in section 4.3.

As discussed in section 4.4.2, the easiest way to interpret the estimated standard deviations of the random intercept and random slope (conditional on the covariates) is to form intervals within which 95% of the schools' random intercepts and slopes are expected to lie assuming normality. Remember that these intervals represent ranges within which 95% of the realizations of a *random variable* are expected to lie, a concept different from confidence intervals, which are ranges within which an *unknown parameter* is believed to lie.

For the intercepts, we obtain $-0.115 \pm 1.96 \times 3.007$, so 95% of schools have their intercept in the range -6.0 to 5.8. In other words, the school mean GCSE scores for children with average LRT scores (`lrt`=0) vary between -6.0 and 5.8. For the slopes, we obtain $0.557 \pm 1.96 \times 0.121$, giving an interval from 0.32 to 0.80. Thus, 95% of schools have slopes between 0.32 and 0.80.

This exercise of forming intervals is particularly important for slopes because it is useful to know whether the slopes have different signs for different schools (which would be odd in the current example). The range from 0.32 to 0.80 is fairly wide and the regression lines for schools may cross: one school could add more value (produce higher mean GCSE scores for given LRT scores) than another school for students with low LRT scores and add less value than the other school for students with high LRT scores.

The estimated correlation $\widehat{\rho}_{21} = 0.50$ between random intercepts and slopes (given the covariates) means that schools with larger mean GCSE scores for students with average LRT scores than other schools also tend to have larger slopes than those other schools. This correlation, combined with the random-intercept and slope variances and the range of LRT scores, determines how much the lines cross, something that is best explored by plotting the predicted regression lines for the schools, as demonstrated in section 4.8.3.

The variance of the total residual ξ_{ij} (equal to the conditional variance of the responses y_{ij} given the covariates $\mathbf{X}_j$) was given in (4.2). We can estimate the corresponding standard deviation by plugging in the ML estimates:

$$\begin{aligned}\sqrt{\widehat{\mathrm{Var}}(\xi_{ij}|\mathbf{X}_j)} &= \sqrt{\widehat{\psi}_{11} + 2\widehat{\psi}_{21}x_{ij} + \widehat{\psi}_{22}x_{ij}^2 + \widehat{\theta}} \\ &= \sqrt{9.0447 + 2 \times 0.1804 \times x_{ij} + 0.0145 \times x_{ij}^2 + 55.3653}\end{aligned}$$

A graph of the estimated standard deviation of the total residual against the covariate `lrt` (x_{ij}) can be obtained using the following `twoway function` command, which is graphed in figure 4.8:

```
. twoway function sqrt(9.0447+2*0.1804*x+0.0145*x^2+55.3653), range(-30 30)
> xtitle(LRT) ytitle(Estimated standard deviation of total residual)
```

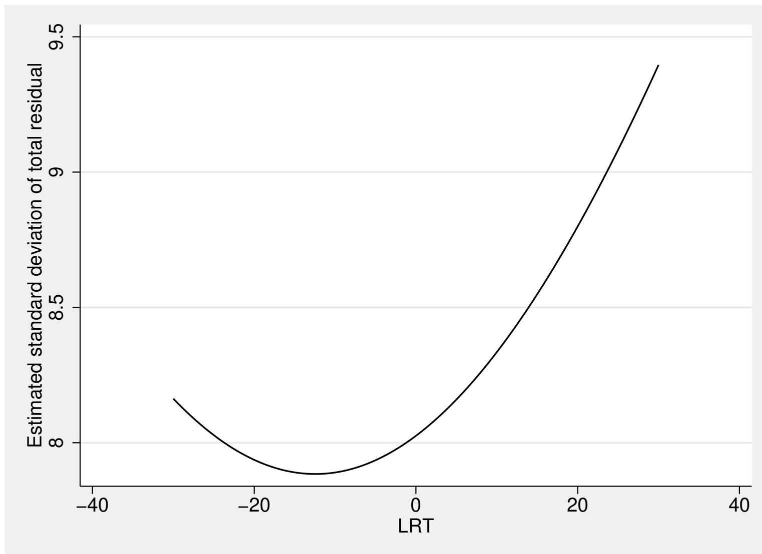

Figure 4.8: Heteroskedasticity of total residual ξ_{ij} as function of `lrt`

The estimated standard deviation of the total residual varies between just under 8 and just under 9.5 for the range of `lrt` in the data.

4.8 Assigning values to the random intercepts and slopes

Having obtained estimated model parameters $\widehat{\beta}_1$, $\widehat{\beta}_2$, $\widehat{\psi}_{11}$, $\widehat{\psi}_{22}$, $\widehat{\psi}_{21}$, and $\widehat{\theta}$, we now assign values to the random intercepts and slopes (see also section 2.11). This is useful for model visualization, residual diagnostics, and inference for individual clusters, as will be demonstrated. Until section 4.8.5, the estimated parameters will be treated as known. In section 4.8.5, we will use REML estimation to obtain standard errors for empirical Bayes predictions that take uncertainty in estimating $\widehat{\beta}_1$ and $\widehat{\beta}_2$ into account.

4.8.1 Maximum "likelihood" estimation

Maximum "likelihood" estimates of the random intercepts and slopes can be obtained by first predicting the total residuals $\widehat{\xi}_{ij} = y_{ij} - (\widehat{\beta}_1 + \widehat{\beta}_2 x_{ij})$ and then fitting individual regressions of $\widehat{\xi}_{ij}$ on x_{ij} for each school by OLS. As explained in section 2.11.1, we put "likelihood" in quotes in the section heading because it differs from the marginal likelihood that is used to estimate the model parameters.

We can fit the individual regression models by using the statsby prefix command as shown in section 4.3. We first retrieve the mixed estimates stored under rc,

```
. estimates restore rc
(results rc are active now)
```

and obtain the predicted total residuals,

```
. predict fixed, xb
. generate totres = gcse - fixed
```

We can then use statsby to produce the variables mli and mls, which contain the ML estimates $\widehat{\zeta}_{1j}$ and $\widehat{\zeta}_{2j}$ of the random intercepts and slopes, respectively:

```
. statsby mli=_b[_cons] mls=_b[lrt], by(school) saving(ols, replace):
> regress totres lrt
(running regress on estimation sample)
      Command: regress totres lrt
          mli: _b[_cons]
          mls: _b[lrt]
           By: school
Statsby groups
────┼─── 1 ───┼─── 2 ───┼─── 3 ───┼─── 4 ───┼─── 5
..................................................    50
...............
. sort school
. merge m:1 school using ols
    Result                      Number of obs
    ─────────────────────────────────────────
    Not matched                             0
    Matched                             4,059  (_merge==3)
    ─────────────────────────────────────────
. drop _merge
```

Maximum likelihood estimates will not be available for schools with only one observation or for schools within which x_{ij} does not vary. There are no such schools in the dataset, but school 48 has only two observations, and the ML estimates of the intercept and slope look odd:

```
. list lrt gcse mli mls if school==48, clean noobs
        lrt      gcse       mli         mls
    -4.5541   -1.2908   -32.607   -7.458484
    -3.7276   -6.9951   -32.607   -7.458484
```

Because there are only two students, the fitted line connects the points perfectly. The school's intercept and slope are determined by ϵ_{1j} and ϵ_{2j} roughly as much as they are by the true intercept and slope. The intercept and slope estimates are therefore imprecise and can be extreme; the so-called "bouncing beta" phenomenon often encountered when using ML estimation of random effects for clusters that provide little information. In general, we therefore do not recommend using this method and suggest using empirical Bayes prediction instead.

4.8.2 Empirical Bayes prediction

As discussed for random-intercept models in section 2.11.2, empirical Bayes (EB) predictions have a smaller prediction error variance (for given model parameters) than ML estimates because of shrinkage toward the mean. Furthermore, EB predictions are available for schools with only one observation or only one unique value of x_{ij}, for which ML estimates cannot be obtained.

Empirical Bayes predictions $\widetilde{\zeta}_{1j}$ and $\widetilde{\zeta}_{2j}$ of the random intercepts ζ_{1j} and slopes ζ_{2j}, respectively, can be obtained using the `predict` command with the `reffects` option after estimation with `mixed`:

```
. estimates restore rc
. predict ebs ebi, reffects
```

Here we specified the variable names `ebs` and `ebi` for the EB predictions $\widetilde{\zeta}_{2j}$ and $\widetilde{\zeta}_{1j}$ of the random slopes and intercepts. The intercept variable comes last because `mixed` treats the intercept as the last random effect, as reflected by the output. This order is consistent with Stata's convention of treating the fixed intercept as the last regression parameter in estimation commands.

To compare the EB predictions with the ML estimates, we list one observation per school for schools 1–9 and school 48:

```
. list school mli ebi mls ebs if pickone==1 & (school<10 | school==48), noobs
```

school	mli	ebi	mls	ebs
1	3.948387	3.749336	.1526116	.1249761
2	4.937838	4.702127	.2045585	.1647271
3	5.69259	4.797687	.0222565	.0808662
4	.1526221	.3502472	.2047174	.1271837
5	2.719525	2.462807	.1232876	.0720581
6	6.147151	5.183819	-.0213858	.0586235
7	4.100312	3.640948	-.314454	-.1488728
8	-.136885	-.1218853	.0106781	.0068856
9	-2.258599	-1.767985	-.1555332	-.0886202
48	-32.607	-.4098203	-7.458484	-.0064854

Most of the time, the EB predictions are closer to 0 than the ML estimates because of shrinkage, as discussed for random-intercept models in section 2.11.2. However, for models with several random effects, the relationship between EB predictions and ML estimates is somewhat more complex than for random-intercept models. The benefit of shrinkage is apparent for school 48, where the EB predictions appear more reasonable than the ML estimates.

We can see shrinkage more clearly by plotting the EB predictions against the ML estimates and superimposing a $y = x$ line. For the random intercept, the command is

```
. twoway (scatter ebi mli if pickone==1 & school!=48, mlabel(school))
> (function y=x, range(-10 10)), xtitle(ML estimate)
> ytitle(EB prediction) legend(off) xline(0)
```

and for the random slope, it is

```
. twoway (scatter ebs mls if pickone==1 & school!=48, mlabel(school))
> (function y=x, range(-0.6 0.6)), xtitle(ML estimate)
> ytitle(EB prediction) legend(off) xline(0)
```

These commands produce the graphs in figure 4.9 (we excluded school 48 from the graphs because the ML estimates are so extreme).

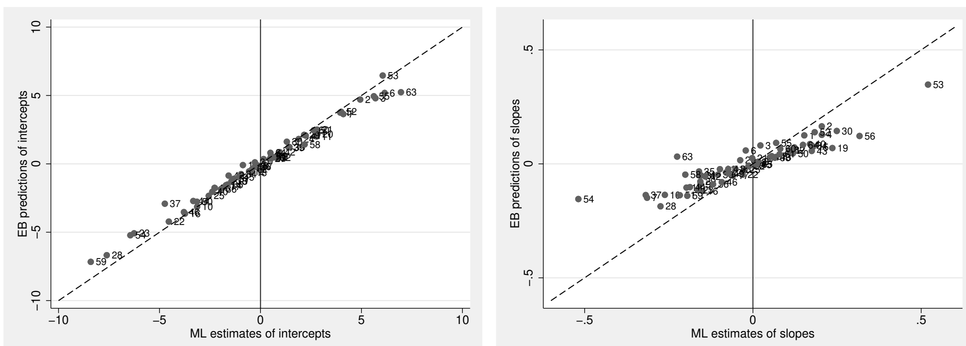

Figure 4.9: Scatterplots of empirical Bayes (EB) predictions versus maximum likelihood (ML) estimates of school-specific intercepts (left) and slopes (right); equality of EB and ML shown as dashed reference lines and ML estimates of 0 shown as solid reference lines

For ML estimates above 0, the EB prediction tends to be smaller than the ML estimate; the reverse is true for ML estimates below 0. There is more shrinkage for slopes than for intercepts.

4.8.3 Model visualization

To better understand the estimates obtained for random-intercept models and random-coefficient models—and in particular, the variability implied by the random part—it is useful to produce graphs of predicted model-implied regression lines for the individual schools.

This can be achieved using the `predict` command with the `fitted` option to obtain school-specific fitted regression lines, with ML estimates substituted for the regression parameters (β_1 and β_2) and EB predictions substituted for the random effects (ζ_{1j} for the random-intercept model, and ζ_{1j} and ζ_{2j} for the random-coefficient model). For instance, for the random-coefficient model, the predicted regression line for school j is

$$\widehat{y}_{ij} \;=\; \widehat{\beta}_1 + \widehat{\beta}_2 x_{ij} + \widetilde{\zeta}_{1j} + \widetilde{\zeta}_{2j} x_{ij}$$

These predictions are obtained by typing

```
. predict murc, fitted
```

and a spaghetti plot is produced as follows:

```
. sort school lrt
. twoway (line murc lrt, connect(ascending)), xtitle(LRT)
> ytitle(Empirical Bayes regression lines for model 2)
```

To obtain predictions for the random-intercept model, we must first restore the estimates stored under the name `ri`:

```
. estimates restore ri
(results ri are active now)
. predict muri, fitted
. sort school lrt
. twoway (line muri lrt, connect(ascending)), xtitle(LRT)
> ytitle(Empirical Bayes regression lines for model 1)
```

The resulting spaghetti plots of the school-specific regression lines for both the random-intercept model and the random-coefficient model are given in figure 4.10.

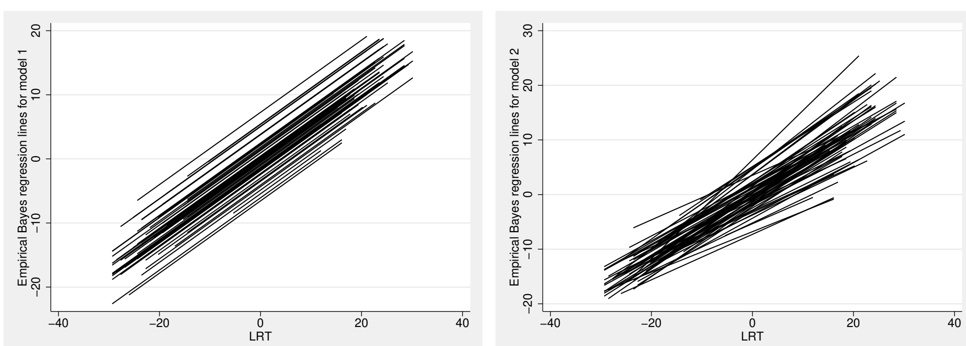

Figure 4.10: Spaghetti plots of empirical Bayes (EB) predictions of school-specific regression lines for the random-intercept model (left) and the random-coefficient model (right)

The predicted school-specific regression lines are parallel for the random-intercept model (with vertical shifts from the population-averaged regression line given by the $\widetilde{\zeta}_{1j}$) but are not parallel for the random-coefficient model, where the slopes $\beta_2 + \widetilde{\zeta}_{2j}$ also vary across schools. Because of shrinkage, the predicted lines vary somewhat less than implied by the estimated variances and covariance.

When there are many clusters, spaghetti plots become messy, and it may be a good idea to plot the lines for a random sample of clusters (see *Part III: Introduction to models for longitudinal and panel data* for an example).

4.8.4 Residual diagnostics

If normality is assumed for the random intercepts ζ_{1j}, random slopes ζ_{2j}, and level-1 residuals ϵ_{ij}, the corresponding EB predictions should also have normal distributions.

To plot the distributions of the predicted random effects, we must pick one prediction per school, and we can accomplish this by using the `pickone` variable created with the command `egen pickone = tag(school)` in section 4.3. We can now plot the distributions by using

```
. histogram ebi if pickone==1, normal xtitle(Predicted random intercepts)
. histogram ebs if pickone==1, normal xtitle(Predicted random slopes)
```

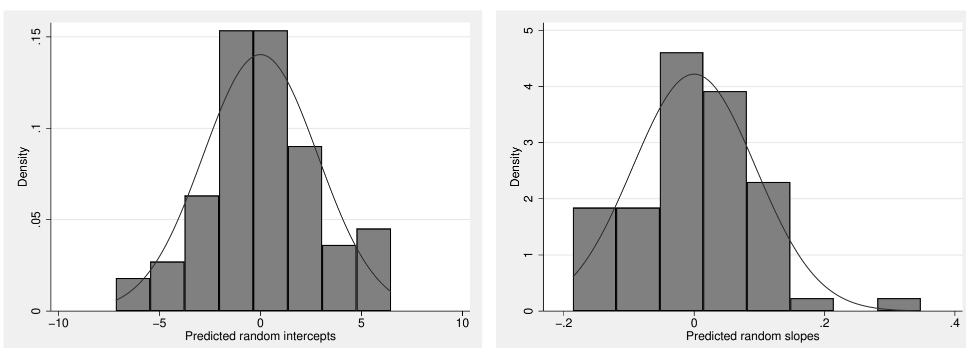

Figure 4.11: Histograms of predicted random intercepts and slopes

The histograms in figure 4.11 look approximately normal although the one for the slopes is perhaps a little positively skewed. It should be noted, however, that moderate nonnormality of random effects can easily be missed because EB predictions tend to be closer to normal than the true random effects.

It is also useful to look at the bivariate distribution of the predicted random intercepts and slopes by using a scatterplot, or to display such a scatterplot together with the two histograms:

```
. scatter ebs ebi if pickone==1, saving(yx, replace)
> xtitle("Random intercept") ytitle("Random slope") ylabel(, nogrid)
. histogram ebs if pickone==1, freq horizontal saving(hy, replace) normal
> yscale(alt) ytitle(" ") fxsize(35) ylabel(, nogrid)
. histogram ebi if pickone==1, freq saving(hx, replace) normal
> xscale(alt) xtitle(" ") fysize(35) ylabel(, nogrid)
. graph combine hx.gph yx.gph hy.gph, hole(2) imargin(0 0 0 0)
```

Here the scatterplot and histograms are first plotted separately and then combined using the `graph combine` command. In the first `histogram` command, the `horizontal` option is used to produce a rotated histogram of the random slopes. In the histogram commands, the `yscale(alt)` and `xscale(alt)` options are used to put the corresponding axes on the other side, and the `normal` option is used to overlay normal density curves. The `fysize(35)` and `fxsize(35)` options change the aspect ratios of the histograms, making them more flat so that they use up a smaller portion of the combined graph. Finally, in the `graph combine` command, the graphs are listed in lexicographic order, the `hole(2)` option denotes that there should be a hole in the second position—that is, the top-right corner—and the `imargin(0 0 0 0)` option reduces the space between the graphs. The resulting graph is shown in figure 4.12.

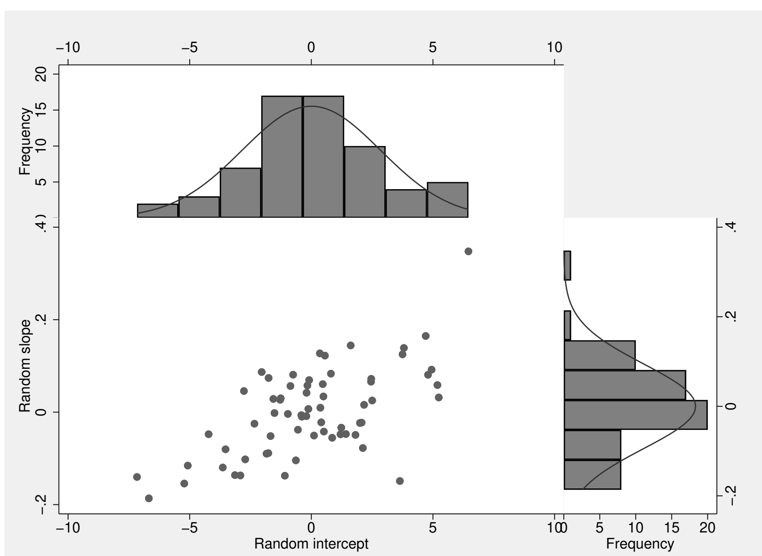

Figure 4.12: Scatterplot and histograms of predicted random intercepts and slopes

After estimation with `mixed`, we obtain the predicted level-1 residuals,

$$\widetilde{\epsilon}_{ij} \;=\; y_{ij} - (\widehat{\beta}_1 + \widehat{\beta}_2 x_{ij} + \widetilde{\zeta}_{1j} + \widetilde{\zeta}_{2j} x_{ij})$$

by using

```
. predict res1, residuals
```

We plot the residuals by using the following command, which produces the graph in figure 4.13:

```
. histogram res1, normal xtitle(Predicted level-1 residuals)
```

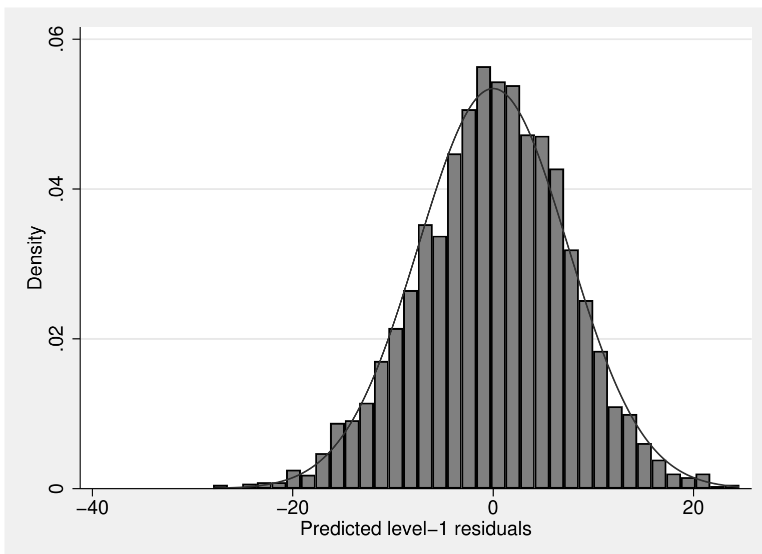

Figure 4.13: Histogram of predicted level-1 residuals

To obtain standardized level-1 residuals, use the `rstandard` option in the `predict` command after estimation using `mixed`.

4.8.5 Inferences for individual schools

Random-intercept predictions $\widetilde{\zeta}_{1j}$ are sometimes viewed as measures of institutional performance—in the present context, how much value the schools add for children with LRT scores equal to 0 (the mean). However, we may not have adequately controlled for covariates correlated with achievement that are outside the control of the school, such as student SES. Furthermore, the model assumes that the random intercepts are uncorrelated with the LRT scores, so if schools with higher mean LRT scores tend to add more value, their value added would be underestimated. Nevertheless, predicted random intercepts shed some light on the research question: Which schools are most effective for children with LRT = 0?

It does not matter whether we add the predicted fixed part of the model because the ranking of schools is not affected by this.

Returning to the question of comparing the schools' effectiveness for children with LRT scores equal to 0, we can plot the predicted random intercepts with approximate 95% confidence intervals based on comparative standard errors (see section 2.11.3). We recommend fitting the model by REML before using `predict` in order to obtain estimated comparative standard errors that take uncertainty in the estimated regression coefficients into account:

```
. quietly mixed gcse lrt || school: lrt, covariance(unstructured) reml
. predict slope1 inter1, reffects reses(slope_se inter_se)
```

Here we only need `inter_se`. We first produce ranks for the schools in ascending order of the random-intercept predictions `inter1`:

```
. gsort + inter1 - pickone
. generate rank = sum(pickone)
```

Here the `gsort` command is used to sort in ascending order of `inter1` (indicated by "`+ inter1`") and, within `inter1`, in descending order of `pickone` (indicated by "`-pickone`"). The `sum()` function forms the cumulative sum, so the variable `rank` increases by 1 every time a new school with higher value of `inter1` is encountered. Before producing the graph, we generate a variable, `labpos`, for the vertical positions in the graph where the school identifiers should go:

```
. generate labpos = inter1 + 1.96*inter_se + .5
```

We are now ready to produce a so-called *caterpillar plot*:

```
. serrbar inter1 inter_se rank if pickone==1, addplot(scatter labpos rank,
> mlabel(school) msymbol(none) mlabpos(0)) scale(1.96) xtitle(Rank)
> ytitle(Prediction) legend(off)
```

The school labels were added to the graph by superimposing a scatterplot onto the error bar plot with the `addplot()` option, where the vertical positions of the labels are given by the variable `labpos`. The resulting caterpillar plot is shown in figure 4.14.

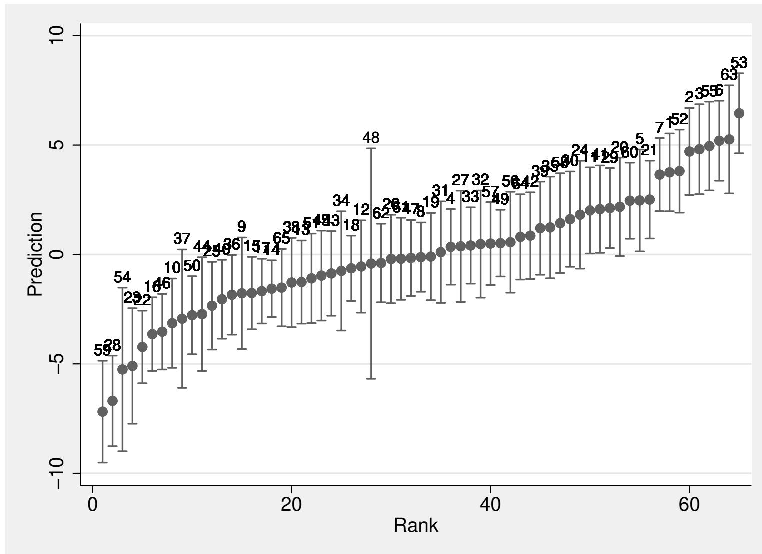

Figure 4.14: Caterpillar plot of random-intercept predictions and approximate 95% confidence intervals versus ranking (school identifiers shown on top of confidence intervals)

The interval for school 48 is particularly wide because there are only two students from this school in the dataset. It is clear from the large confidence intervals that the rankings are not precise and that perhaps only a coarse classification into poor, medium, and good schools can be justified.

An alternative method for producing a caterpillar plot is to first generate the confidence limits `lower` and `upper`,

```
. generate lower = inter1 - 1.96*inter_se
. generate upper = inter1 + 1.96*inter_se
```

and then use the `rcap` plot type to produce the intervals:

```
. twoway (rcap lower upper rank, blpatt(solid) lcol(black))
> (scatter inter1 rank)
> (scatter labpos rank, mlabel(school) msymbol(none) mlabpos(0)
> mlabcol(black) mlabsiz(medium)),
> xtitle(Rank) ytitle(Prediction) legend(off)
> xscale(range(1 65)) xlabel(1/65) ysize(1)
```

Here `scatter` is first used to overlay the point estimates and then the labels. The `ysize()` option is used to change the aspect ratio and obtain the horizontally stretched graph shown in figure 4.15.

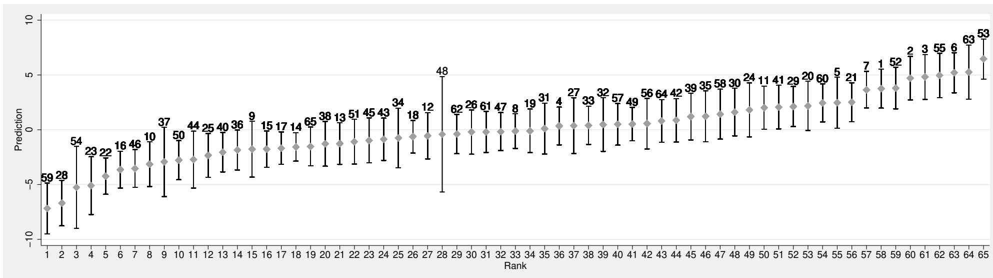

Figure 4.15: Stretched caterpillar plot of random-intercept predictions and approximate 95% confidence intervals versus ranking (school identifiers shown on top of confidence intervals)

We could also produce similar plots for children with different values x^0 of the LRT scores:

$$\widehat{\beta}_1 + \widehat{\beta}_2 x^0 + \widetilde{\zeta}_{1j} + \widetilde{\zeta}_{2j} x^0$$

For instance, in a similar application, Goldstein et al. (2000) substitute the 10th percentile of the intake measure to compare school effectiveness for poorly performing children. (To obtain confidence intervals for different values of x^0 requires posterior correlations that can be obtained by `gllamm`; see the `gllamm` companion.)

4.9 Two-stage model formulation

In this section, we describe an alternative way of specifying random-coefficient models that is popular in some areas such as education (for example, Raudenbush and Bryk 2002). As shown below, models are specified in two stages (for levels 1 and 2), necessitating a distinction between level-1 and level-2 covariates. Many people find this formulation helpful for interpreting and specifying models. Identical models can be formulated using either the approach discussed up to this point or the two-stage formulation.

To express the random-coefficient model by using a two-stage formulation, Raudenbush and Bryk (2002) specify a level-1 model:

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + r_{ij}$$

where the intercept β_{0j} and slope β_{1j} are school-specific coefficients. Their level-2 models have these coefficients as responses:

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j} \end{aligned} \qquad (4.4)$$

Sometimes the first of these level-2 models is referred to as a "means as outcomes" or "intercepts as outcomes" model, and the second as a "slopes as outcomes" model. It is typically assumed that given the covariate(s), the residuals or disturbances u_{0j} and u_{1j} in the level-2 model have a bivariate normal distribution with 0 mean and covariance matrix

$$\mathbf{T} \;=\; \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix}, \qquad \tau_{10} = \tau_{01}$$

The level-2 models cannot be fit on their own because the school-specific coefficients β_{0j} and β_{1j} are not observed. Instead, we must substitute the level-2 models into the level-1 model to obtain the *reduced-form* model for the observed responses, y_{ij}:

$$\begin{aligned} y_{ij} &= \underbrace{\gamma_{00} + u_{0j}}_{\beta_{0j}} + \underbrace{(\gamma_{10} + u_{1j})}_{\beta_{1j}} x_{ij} + r_{ij} \\ &= \underbrace{\gamma_{00} + \gamma_{10}x_{ij}}_{\text{fixed}} + \underbrace{u_{0j} + u_{1j}x_{ij} + r_{ij}}_{\text{random}} \\ &\equiv \beta_1 + \beta_2 x_{ij} \quad + \zeta_{1j} + \zeta_{2j}x_{ij} + \epsilon_{ij} \end{aligned}$$

In the reduced form, the fixed part is usually written first, followed by the random part. As shown in the last line of the equation above, we can return to our previous notation by defining $\beta_1 \equiv \gamma_{00}$, $\beta_2 \equiv \gamma_{10}$, $\zeta_{1j} \equiv u_{0j}$, $\zeta_{2j} \equiv u_{1j}$, and $\epsilon_{ij} \equiv r_{ij}$. The above model is thus equivalent to the model in (4.1).

The level-1 model contains only level-1 covariates (that vary between units within clusters). Any level-2 covariates (that do not vary within clusters) are included in the level-2 models. For instance, we could include dummy variables for type of school: w_{1j} for boys-only schools and w_{2j} for girls-only schools, with mixed schools as the reference category. If we include these dummy variables in the model for the random intercept,

$$\beta_{0j} \;=\; \gamma_{00} + \gamma_{01}w_{1j} + \gamma_{02}w_{2j} + u_{0j}$$

the reduced form becomes

$$\begin{aligned} y_{ij} &= \underbrace{\gamma_{00} + \gamma_{01}w_{1j} + \gamma_{02}w_{2j} + u_{0j}}_{\beta_{0j}} + \underbrace{(\gamma_{10} + u_{1j})}_{\beta_{1j}} x_{ij} + r_{ij} \\ &= \underbrace{\gamma_{00} + \gamma_{01}w_{1j} + \gamma_{02}w_{2j} + \gamma_{10}x_{ij}}_{\text{fixed}} + \underbrace{u_{0j} + u_{1j}x_{ij} + r_{ij}}_{\text{random}} \end{aligned}$$

If we also include the dummy variables for type of school in the model for the random slope,

$$\beta_{1j} \;=\; \gamma_{10} + \gamma_{11}w_{1j} + \gamma_{12}w_{2j} + u_{1j}$$

we obtain so-called *cross-level interactions* between covariates varying at different levels—w_{1j} by x_{ij} as well as w_{2j} by x_{ij}—in the reduced form

$$\begin{aligned} y_{ij} &= \underbrace{\gamma_{00} + \gamma_{01}w_{1j} + \gamma_{02}w_{2j} + u_{0j}}_{\beta_{0j}} + \underbrace{(\gamma_{21} + \gamma_{22}w_{2j} + \gamma_{23}w_{3j} + u_{1j})}_{\beta_{1j}} x_{ij} + r_{ij} \\ &= \underbrace{\gamma_{00} + \gamma_{01}w_{1j} + \gamma_{02}w_{2j} + \gamma_{10}x_{ij} + \gamma_{11}w_{1j}x_{ij} + \gamma_{12}w_{2j}x_{ij}}_{\text{fixed}} + \underbrace{u_{0j} + u_{1j}x_{ij} + r_{ij}}_{\text{random}} \end{aligned}$$

The effect of `lrt` now depends on the type of school, with γ_{11} representing the additional effect of `lrt` on `gcse` for boys-only schools compared with mixed schools and γ_{12} representing the additional effect for girls-only schools compared with mixed schools.

For estimation in `mixed`, it is necessary to convert the two-stage formulation to the reduced form because the fixed part of the model is specified first, followed by the random part of the model. Using factor variables in `mixed`, the command is

```
. mixed gcse i.schgend##c.lrt || school: lrt, covariance(unstructured) mle
> stddeviations vce(robust)

Mixed-effects regression                        Number of obs     =      4,059
Group variable: school                          Number of groups  =         65
                                                Obs per group:
                                                              min =          2
                                                              avg =       62.4
                                                              max =        198
                                                Wald chi2(5)      =     930.12
Log pseudolikelihood = -13998.825               Prob > chi2       =     0.0000

                                (Std. err. adjusted for 65 clusters in school)
------------------------------------------------------------------------------
             |               Robust
        gcse | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
     schgend |
        boys |   .8546715   .9648313     0.89   0.376    -1.036363    2.745706
       girls |    2.43341    .836635     2.91   0.004     .7936359    4.073185
             |
         lrt |   .5712361   .0235687    24.24   0.000     .5250423      .61743
             |
schgend#c.lrt|
        boys |  -.0230098   .0541404    -0.43   0.671    -.1291229    .0831034
       girls |   -.029544   .0493976    -0.60   0.550    -.1263616    .0672735
             |
       _cons |  -.9976073   .5544969    -1.80   0.072    -2.084401    .0891867
------------------------------------------------------------------------------

------------------------------------------------------------------------------
                             |               Robust
  Random-effects parameters  |   Estimate   std. err.     [95% conf. interval]
-----------------------------+------------------------------------------------
school: Unstructured         |
                     sd(lrt) |   .1199154   .0242665      .0806535    .1782899
                   sd(_cons) |   2.797934   .3028989      2.263018     3.45929
             corr(lrt,_cons) |   .5967727   .1391651      .2584913    .8046796
-----------------------------+------------------------------------------------
                sd(Residual) |   7.441831   .1251251      7.200588    7.691158
------------------------------------------------------------------------------
```

Here mixed schools are the reference category for `schgend` to which boys-only schools and girls-only schools are compared. We see that, when `lrt` is 0, students from girls-only schools perform significantly better at the 5% level than students from mixed schools, whereas students from boys-only schools do not perform significantly better than students from mixed schools. The effect of `lrt` does not differ significantly between boys-only schools and mixed schools or between girls-only schools and mixed schools. The estimates and the corresponding parameters in the two-stage formulation are given under "Rand. coefficient & level-2 covariates" in the last three columns of table 4.1.

Although equivalent models can be specified using either the reduced-form (used by `mixed`) or the two-stage (used by the HLM software of Raudenbush et al. [2019]) formulation, in practice, model specification to some extent depends on the approach adopted. For instance, cross-level interactions are easily included using the two-stage specification in HLM, whereas same-level interactions must be created outside the program. Papers using HLM therefore tend to include more cross-level interactions and fewer same-level interactions. They also tend to include more random coefficients than papers using, for instance, Stata because the level-2 models look odd without residuals.

4.10 Some warnings about random-coefficient models

4.10.1 Meaningful specification

It rarely makes sense to include a random slope if there is no random intercept, just like interactions between two covariates usually do not make sense without including the covariates themselves in standard regression models. Similarly, it is seldom sensible to include a random slope without including the corresponding fixed slope because it is usually strange to allow the slope to vary randomly but constrain its population mean to 0.

It is generally not a good idea to include a random coefficient for a covariate that does not vary at a lower level than the random coefficient itself. For example, in the inner-London-schools data, it does not make sense to include a school-level random slope for type of school because type of school does not vary within schools. Because we cannot estimate the effect of type of school for individual schools, it also appears impossible to estimate the variability of the effect of type of school between schools. However, level-2 random coefficients of level-2 covariates can be used to construct heteroskedastic random intercepts (see section 7.5.2).

4.10.2 Many random coefficients

It may be tempting to allow many different covariates to have random slopes. However, the number of parameters for the random part of the model increases rapidly with the number of random slopes because there is a variance parameter for each random effect (intercept or slope) and a covariance parameter for each *pair* of random effects. If there are k random slopes plus one random intercept, then there are $(k+2)(k+1)/2+1$ parameters in the random part (for example, $k=3$ gives 11 parameters).

Another problem is that clusters may not provide much information on cluster-specific slopes and hence on the slope variance either if the clusters are small, or if x_{ij} does not vary much within clusters or varies only in a small number of clusters. Perhaps a useful rule is to consider the random part of the model (ignoring the fixed part) and replace the random effects with fixed regression coefficients. It should be possible (even if not very sensible) to fit the resulting model to a good number of clusters (say, 20 or more), with some error degrees of freedom. Note, however, that it does not matter if some of the clusters have insufficient data as long as there are an adequate number of clusters that do have sufficient data. It is never a good idea to discard clusters merely because they provide little information on some of the parameters of the model.

In general, it makes sense to allow for more flexibility in the fixed part of the model than in the random part. For instance, the fixed part of the model may include a dummy variable for each occasion in longitudinal data, but in the random part of the model it may be sufficient to allow for a random intercept and a random slope of time, keeping in mind that in this case it is only assumed that the *deviation* from the population-average curve is linear in time, not that the relationship itself is linear. See section 7.3 for examples of modeling a nonlinear relationship in the fixed part of the model but not in the random part.

The overall message is that random slopes should be included only if strongly suggested by the subject-matter theory related to the application *and* if the data provide sufficient information.

4.10.3 Convergence problems

Convergence problems can manifest themselves in different ways. Either estimates are never produced, or standard errors are missing, or `mixed` produces messages such as "nonconcave", or "backed-up", or "standard error calculation has failed". Sometimes none of these things happen, but the confidence intervals for some of the correlations cover the full permissible range from −1 to 1 (see sections 7.3 and 8.13.2 for examples).

Convergence problems can occur because the estimated covariance matrix "tries" to become negative definite, meaning, for instance, that variances try to become negative or correlations try to be greater than 1 or less than −1. All the commands in Stata force the covariance matrix to be positive (semi)definite, and when parameters approach nonpermissible values, convergence can be slow or even fail. It may help to translate and rescale x_{ij} because variances and covariances are not invariant to these transformations. Often a better remedy is to simplify the model by removing some random slopes. Convergence problems can also occur because of lack of identification, and again, a remedy is to simplify the model.

However, before giving up on a model, it is worth attempting to achieve convergence by trying both the `mle` and the `reml` options, specifying the `difficult` option, trying the `matlog` option (which parameterizes the random part differently during maximization), or increasing the number of EM iterations by using either the `emiterate()` option or even the `emonly` option. It can also be helpful to monitor the iterations more closely by using `trace`, which displays the parameter estimates at the end of each iteration (unfortunately, not for the EM iterations). Lack of identification of a parameter might be recognized by that parameter changing wildly between iterations without much of a change in the log likelihood. Problems with a variance approaching 0 can be detected by noticing that the log-standard deviation takes on very large negative values.

4.10.4 Lack of identification

Sometimes random-coefficient models are simply not identified (or in other words, underidentified). As an important example, consider balanced data with clusters of size $n_j = 2$ and with a covariate x_{ij} taking the same two values $t_1 = 0$ and $t_2 = 1$ for each cluster (an example would be the peak-expiratory-flow data from chapter 2). A model with a random intercept, a random slope of x_{ij}, and a level-1 residual, all of which are normally distributed, is not identified in this case. This can be seen by considering the two distinct variances (for $i = 1$ and $i = 2$) and one covariance of the total residuals when $t_1 = 0$ and $t_2 = 1$:

$$
\begin{aligned}
\text{Var}(\xi_{1j}) &= \psi_{11} + \theta \\
\text{Var}(\xi_{2j}) &= \psi_{11} + 2\psi_{21} + \psi_{22} + \theta \\
\text{Cov}(\xi_{1j}, \xi_{2j}) &= \psi_{11} + \psi_{21}
\end{aligned}
$$

The marginal distribution of y_{ij} given the covariates is normal and therefore completely characterized by the fixed part of the model and these three model-implied moments (two variances and a covariance). However, the three moments are determined by four parameters of the random part (ψ_{11}, ψ_{22}, ψ_{21}, and θ), so fitting the model-implied moments to the data would effectively involve solving three equations for four unknowns. The model is therefore not identified. We could identify the model by setting $\theta = 0$, which does not impose any restrictions on the covariance matrix (however, such a constraint is not allowed in `mixed`). The original model becomes identified if the covariate x_{ij}, which has a random slope, varies also between clusters because the model-implied covariance matrix of the total residuals then differs between clusters, yielding more equations to solve for the four parameters.

Still assuming that the random effects and level-1 residual are normally distributed, consider now the case of balanced data with clusters of size $n_j = 3$ and with a covariate x_{ij} taking the same three values t_1, t_2, and t_3 for each cluster. An example would be longitudinal data with three occasions at times t_1, t_2, and t_3. Instead of including a random intercept and a random slope of time, it may be tempting to specify a random-coefficient model with a random intercept and two random coefficients for the dummy variables for occasions two and three. In total, such a model would contain seven (co)variance parameters: six for the three random effects and one for the level-1 residual variance. Because the covariance matrix of the responses for the three occasions given the covariates only has six elements, it is impossible to solve for all unknowns. The same problem would occur when attempting to fit this kind of model for more than three occasions.

4.11 Summary and further reading

In this chapter, we introduced the notion of slopes or regression coefficients varying randomly between clusters in linear models. Linear random-coefficient models are parsimonious representations of situations where each cluster has a separate regression model with its own intercept and slope. The linear random-coefficient model was applied to a cross-sectional study of school effectiveness. Here students were nested in schools, and we considered school-specific regressions.

An important consideration when using random-coefficient models is that the interpretation of the covariance matrix of the random effects depends on the scale and location of the covariates having random slopes. One should thus be careful when interpreting the variance and covariance estimates. We briefly demonstrated a two-stage formulation of random-coefficient models that is popular in some fields. This formulation can be used to specify models that are equivalent to models specified using the reduced-form formulation used in this book.

The utility of empirical Bayes prediction was demonstrated for visualizing the model, making inferences for individual clusters, and for diagnostics. See Skrondal and Rabe-Hesketh (2009) for a detailed discussion of prediction of random effects.

In section 3.7.4, we discussed the problem of level-2 endogeneity in random-intercept models, where the random intercept is correlated with covariates. In random-coefficient models, random slopes can also be correlated with covariates, a problem that is addressed in section 5.5.2 and by Bates et al. (2014).

Introductory books discussing random-coefficient models include Snijders and Bosker (2012, chap. 5), Kreft and de Leeuw (1998, chap. 3), and Raudenbush and Bryk (2002, chap. 2, 4). Papers and chapters with good overviews of much of the material we covered in chapters 2–4 include Snijders (2004), Duncan, Jones, and Moon (1998), and Steenbergen and Jones (2002); a useful list of multilevel terminology is provided by Diez Roux (2002). These papers and chapters are among those collected in Skrondal and Rabe-Hesketh (2010).

The first six exercises are on standard random-coefficient models applied to data from different disciplines, whereas exercises 4.7 and 4.8 use random-coefficient models that correspond to biometrical genetic models for nuclear family data. Random-coefficient models for longitudinal data, often called growth-curve models, are considered in chapter 7. Exercises 7.2, 7.5, 7.6, and 7.7, and parts of the other exercises in that chapter can be viewed as supplementary exercises for the current chapter. Parts of exercises 6.1 and 6.2 are also relevant.

4.12 Exercises

4.1 ❖ Inner-London-schools data

1. Fit the random-coefficient model fit in section 4.9 by explicitly constructing the covariates (not using factor variables).
2. Write down a model with the same covariates as in step 1 that also allows the mean for mixed schools to differ between boys and girls controlling for LRT (`girl` is a dummy for the student being a girl.) Write down null hypotheses in terms of linear combinations of regression coefficients for the following research questions:
   a. Do girls perform better in girls-only schools than in mixed schools (after controlling for the other covariates)?
   b. Do boys perform better in boys-only schools than in mixed schools (after controlling for the other covariates)?
3. Fit the model from step 2, and test the null hypotheses from step 2. Discuss whether there is evidence that children of a given gender perform better in single-sex schools.

4.2 High-school-and-beyond data

Raudenbush and Bryk (2002) and Raudenbush et al. (2019) analyzed data from the High School and Beyond Survey.

The dataset `hsb.dta` has the following variables:

- Level 1 (student)
  - `mathach`: a measure of mathematics achievement
  - `minority`: dummy variable for student being non-White (versus White)
  - `female`: dummy variable for student being female (versus male)
  - `ses`: socioeconomic status (SES) based on parental education, occupation, and income
- Level 2 (school)
  - `schoolid`: school identifier
  - `sector`: dummy variable for school being Catholic (versus public)
  - `pracad`: proportion of students in the academic track
  - `disclim`: scale measuring disciplinary climate
  - `himinty`: dummy variable for more than 40% minority enrollment

Raudenbush et al. (2019) specify a two-level model. We will use their model and notation here. At level 1, math achievement Y_{ij} is regressed on student's SES, centered around the school mean:

$$Y_{ij} \; = \; \beta_{0j} + \beta_{1j}(X_{1ij} - \overline{X}_{1.j}) + r_{ij}, \quad r_{ij} \sim N(0, \sigma^2)$$

where X_{1ij} is the student's SES, $\overline{X}_{1.j}$ is the school mean SES, and r_{ij} is a level-1 residual. At level 2, the intercepts and slopes are regressed on the dummy variable

W_{1j} for the school being a Catholic school (`sector`) and on the school mean SES

$$\beta_{pj} \ = \ \gamma_{p0} + \gamma_{p1} W_{1j} + \gamma_{p2} \overline{X}_{1.j} + u_{pj}, \quad p{=}0,1, \quad (u_{0j}, u_{1j})' \sim N(\mathbf{0}, \mathbf{T})$$

where u_{pj} is a random effect (a random intercept if $p=0$ and a random slope if $p=1$). The covariance matrix

$$\mathbf{T} = \left[\begin{array}{cc} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{array} \right]$$

has three unique elements with $\tau_{10}=\tau_{01}$.

1. Substitute the level-2 models into the level-1 model, and write down the resulting reduced form using the notation of this book.
2. Construct the variables `meanses`, equal to the school-mean SES ($\overline{X}_{1.j}$), and `devses`, equal to the deviations of the student's SES from their school means ($X_{1ij} - \overline{X}_{1.j}$).
3. Fit the model considered by Raudenbush et al. (2019) by using ML in `mixed` and interpret the coefficients. In particular, interpret the estimate of γ_{12}.
4. Fit the model that also includes `disclim` in the level-2 models and `minority` in the level-1 model.

4.3 Homework data

Kreft and de Leeuw (1998) consider a subsample of eighth grade students from the National Education Longitudinal Study of 1988 (NELS–88) collected by the National Center for Educational Statistics of the U.S. Department of Education. The students are viewed as nested in schools.

The data are given in `homework.dta`. In this exercise, we will use the following subset of the variables:

- `schid`: school identifier
- `math`: continuous measure of achievement in mathematics (standardized to have a mean of 50 and a standard deviation of 10)
- `homework`: number of hours of homework done per week
- `white`: student's race (1: White; 0: non-White)
- `ratio`: class size as measured by the student−teacher ratio
- `meanses`: school mean socioeconomic status (SES)

1. Write down and state the assumptions of a random-coefficient model with `math` as response variable and `homework`, `white`, and `ratio` as covariates. Let the intercept and the effect of `homework` vary between schools.
2. Fit the model by REML using Kenward−Roger degrees of freedom and interpret the parameter estimates.
3. Derive an expression for the estimated variance of math achievement conditional on the covariates.

4. How would you extend the model to investigate whether the effect of homework on math achievement depends on the mean SES of schools? Write down both the two-stage and the reduced-form formulation of your extended model.
5. Fit the model from step 4.

4.4 Wheat-and-moisture data

Littell et al. (2006) describe data on ten randomly chosen varieties of winter wheat. Each variety was planted on six randomly selected 1-acre plots of land in a 60-acre field. The amount of moisture in the top 36 inches of soil was determined for each plot before planting the wheat. The response variable is the yield in bushels per acre.

The data, `wheat.dta`, contain the following variables:

- `variety`: variety (or type) of wheat (j)
- `plot`: plot (1 acre) on which wheat was planted (i)
- `yield`: yield in bushels per acre (y_{ij})
- `moist`: amount of moisture in top 36 inches of soil prior to planting (x_{ij})

In this exercise, variety of wheat will be treated as the cluster.

1. Write down the model for `yield` with a fixed and random intercept for variety of wheat and a fixed and variety-specific random slope of `moist`. State all model assumptions.
2. Fit the random-coefficient model by using REML with Kenward–Roger degrees of freedom.
3. Use a likelihood-ratio test to test the null hypothesis that the random-coefficient variance is 0 (although this asymptotic test may require a larger number of clusters).
4. For the chosen model, obtain the predicted yields for each variety (with EB predictions substituted for the random effects).
5. Produce a trellis graph of predicted yield versus moisture, using the `by()` option to obtain a separate graph for each variety.
6. Produce the same graphs as above but with observed values of yield added as dots.

4.5 Well-being-in-the-U.S.-army data Solutions

Bliese (2009) provides the data analyzed by Bliese and Halverson (1996). The data are on soldiers (with the lowest five enlisted ranks) from 99 U.S. army companies in noncombat environments stationed in the U.S. and Europe.

The variables in the dataset `army.dta` are the following:

- `grp`: army company identification number
- `wbeing`: well-being assessed using the General Well-Being Schedule (Dupuy 1978), an 18-item scale measuring depression, anxiety, somatic complaints, positive well-being, and emotional control
- `hrs`: answer to the question "How many hours do you usually work in a day?"
- `cohes`: score on horizontal cohesion scale consisting of eight items, including "My closest relationships are with people I work with"
- `lead`: score on an 11-item leadership consideration (vertical cohesion) scale with a typical item being "The noncommissioned officer in this company would lead well in combat"

1. Fit a random-intercept model for `wbeing` with fixed coefficients for `hrs`, `cohes`, and `lead`, and a random intercept for `grp`. Use ML estimation with robust standard errors.
2. Form the cluster means of the three covariates from step 1, and add them as further covariates to the random-intercept model. Which of the cluster means have coefficients that are significant at the 5% level?
3. Refit the model from step 2 after removing the cluster means that have nonsignificant coefficient estimates at the 5% level. Interpret the remaining coefficients and obtain the estimated intraclass correlation.
4. We have included soldier-specific covariates x_{ij} in addition to the cluster means $\overline{x}_{\cdot j}$. The coefficients of the cluster means represent the contextual effects (see section 3.7.6). Use `lincom` to estimate the corresponding between effects.
5. Add a random slope for `lead` to the model in step 3, and compare this model with the model from step 3 using a likelihood-ratio test (Hint: use `lrtest` with the `force` option).
6. Add a random slope for `cohes` to the model chosen in step 5, and compare this model with the model from step 3 using a likelihood-ratio test. Retain the preferred model.
7. Perform residual diagnostics for the level-1 errors, random intercept, and random slope(s). Do the model assumptions appear to be satisfied?

4.6 Dialyzer data

Vonesh and Chinchilli (1997) analyzed data on low-flux dialyzers used to treat patients with end-stage renal disease (kidney disease) to remove excess fluid and waste from their blood. In low-flux hemodialysis, the ultrafiltration rate at which fluid is removed (volume per time) is thought to follow a straight-line relationship with the transmembrane pressure applied across the dialyzer membrane. In a study to investigate this relationship, three centers measured the ultrafiltration rate at several transmembrane pressures for each of several dialyzers, or patients.

The variables in `dialyzer.dta` are as follows:

- `subject`: subject (or dialyzer) identifier
- `tmp`: transmembrane pressure (mmHg)
- `ufr`: ultrafiltration rate (ml/hr)
- `center`: center at which study was conducted

1. For each center, plot a graph of `ufr` versus `tmp` with separate lines for each subject. You may want to use the `by(center)` option.
2. Write down a model that assumes a linear relationship between `ufr` and `tmp` (denoted y_{ij} and x_{ij}, respectively), with mean intercepts and mean slopes differing between the three centers. In the random part of the model, include a random intercept and a random slope of x_{ij}.
3. Fit the model by REML using Kenward–Roger degrees of freedom.
4. Test whether the mean slopes differ significantly at the 5% level for each pair of centers.
5. Plot the estimated mean line for each center on one graph, using `twoway function`.
6. For center 1, produce a trellis graph of the data and fitted subject-specific regression lines.

4.7 ❖ Family-birthweight data Solutions

Rabe-Hesketh, Skrondal, and Gjessing (2008) analyzed a random subset of the birthweight data from the Medical Birth Registry of Norway described in Magnus et al. (2001). There are 1,000 nuclear families each comprising mother, father, and one child (not necessarily the only child in the family).

The data are given in `family.dta`. In this exercise, we will use the following variables:

- `family`: family identifier (j)
- `member`: family member (i) (1: mother; 2: father; 3: child)
- `bwt`: birthweight in grams (y_{ij})
- `male`: dummy variable for being male (x_{1ij})
- `first`: dummy variable for being the first child (x_{2ij})
- `midage`: dummy variable for mother of family member being aged 20–35 at time of birth (x_{3ij})
- `highage`: dummy variable for mother of family member being older than 35 at time of birth (x_{4ij})
- `birthyr`: year of birth minus 1967 (1967 was the earliest birth year in the birth registry) (x_{5ij})

In this dataset, family members are nested within families. Because of additive genetic and environmental influences, there will be a particular covariance structure between the members of the same family. Rabe-Hesketh, Skrondal, and Gjessing (2008) show that the following random-coefficient model can be used to induce the required covariance structure (see also exercise 4.8):

$$y_{ij} = \beta_1 + \zeta_{1j}(M_i + K_i/2) + \zeta_{2j}(F_i + K_i/2) + \zeta_{3j}(K_i/\sqrt{2}) + \epsilon_{ij} \qquad (4.5)$$

where M_i is a dummy variable for mothers, F_i is a dummy variable for fathers, and K_i is a dummy variable for children. The random coefficients ζ_{1j}, ζ_{2j}, and ζ_{3j} are constrained to have the same variance ψ and to be uncorrelated with each other. We assume that $\zeta_{1j} \sim N(0,\psi)$, $\zeta_{2j} \sim N(0,\psi)$, $\zeta_{3j} \sim N(0,\psi)$, and $\epsilon_{ij} \sim N(0,\theta)$. The variances ψ and θ can be interpreted as additive genetic and environmental variances, respectively, and the total residual variance is $\psi + \theta$.

1. Produce the required dummy variables M_i, F_i, and K_i.
2. Generate variables equal to the terms in parentheses in (4.5).
3. Which of the covariance structures available in `mixed` should be specified for the random coefficients (see the help file for details on the `covariance()` option)?
4. Fit the model given in (4.5) by using ML. The model does not include a random intercept, so use the `noconstant` option.
5. Obtain the estimated proportion of the total variance that is attributable to additive genetic effects.
6. Now fit the model including all the covariates listed above and having the same random part as the model in step 3.
7. Interpret the estimated coefficients from step 6.
8. Conditional on the covariates, what proportion of the residual variance is estimated to be due to additive genetic effects?

4.8 ❖ Covariance-structure-for-nuclear-family data

This exercise concerns family data such as those of exercise 4.7 consisting of a mother, father, and child. Here we consider three types of influences on birthweight: additive genetic effects (due to shared genes), common environmental effects (due to shared environment), and unique environmental effects. These random effects have variances σ_A^2, σ_C^2, and σ_E^2, respectively.

The additive genetic effects have the following properties:

- The parents share no genes by descent, so their additive genetic effects are uncorrelated.
- The child shares half its genes with each parent by decent, giving a correlation of 1/2 with each parent.
- The additive genetic variance should be the same for each family member.

For birth outcomes, no two family members share a common environment because they all developed in different wombs. We therefore cannot distinguish between common and unique environmental effects.

Rabe-Hesketh, Skrondal, and Gjessing (2008) show that we can use the following random-coefficient model to produce the required covariance structure:

$$y_{ij} \;=\; \beta_1 + \zeta_{1j}(M_i + K_i/2) + \zeta_{2j}(F_i + K_i/2) + \zeta_{3j}(K_i/\sqrt{2}) + \epsilon_{ij} \qquad (4.6)$$

where M_i, F_i, and K_i are dummy variables for mothers, fathers, and children, respectively. The random coefficients ζ_{1j}, ζ_{2j}, and ζ_{3j} produce the required additive genetic correlations and variances. These random coefficients are constrained to have the same variance $\psi = \sigma_A^2$ and to be uncorrelated with each other. We assume that $\zeta_{1j} \sim N(0, \sigma_A^2)$, $\zeta_{2j} \sim N(0, \sigma_A^2)$, $\zeta_{3j} \sim N(0, \sigma_A^2)$, and $\epsilon_{ij} \sim N(0, \theta)$.

1. By substituting the appropriate numerical values for the dummy variables M_i, F_i, and K_i in (4.6), write down three separate models, one for mothers, one for fathers, and one for children. It is useful to substitute $i = 1$ for mothers, $i = 2$ for fathers, and $i = 3$ for children in these equations.
2. Using the equations from step 1, demonstrate that the total variance is the same for mothers, fathers, and children.
3. Using the equations from step 1, demonstrate that the covariance between mothers and fathers from the same families is 0.
4. Using the equations from step 1, demonstrate that the correlation between the additive genetic components (terms involving ζ_{1j}, ζ_{2j}, or ζ_{3j}) of mothers and their children is 1/2.
5. What is the interpretation of θ in terms of variances of the common and unique environment effects, σ_C^2 and σ_E^2, respectively?

4.9 ❖ Effect of covariate translation on random-effects covariance matrix

Using (4.2) and the estimates for the random-coefficient model without level-2 covariates given in section 4.5.2, calculate what values $\widehat{\psi}_{11}$ and $\widehat{\psi}_{21}$ would take if you were to subtract 5 from the variable `lrt` and refit the model.

(Pages omitted)