

A First Look at Multilevel Models
Institute for Social Research
Statistical Consulting Service
October–November , 2001
York University

Georges Monette*
with
Qing Shao and Ernest Kwan

May 24, 2002

Contents

1	Introduction	3
2	Preliminaries	6
2.1	Is β_1 the effect of X_1 ?	6
2.2	Visualizing multivariate variance	7
2.2.1	Geometry of regression	9
2.3	Combining estimators	9
2.3.1	Univariate estimators	9

*Many colleagues contributed ideas and data to the material in these notes. In mentioning Pauline Wong, Tammy Kostecki-Dillon, Yifaht Korman, John Fox, Michael Friendly, Qing Shao, Ernest Kwan, Barry Smith, Rona Atlas, Joanne Cummings, H el ene Lavertu, Katherine Henderson, Vincent Hildebrand and Joy Andres-Lemay, I am omitting many other worthy contributors.

†These notes are still very rough and I would appreciate it they were not formally quoted without permission.

2.3.2	Multivariate estimators	9
2.4	Simpson and Robinson	9
2.5	Matrix formulation of regression	10
3	The Multilevel Model	12
3.1	Within school model	13
3.2	Model to compare two schools	13
3.3	Within school model for many schools	14
3.4	Between school model	15
4	Combined model	17
4.1	Multilevel form	17
4.2	Matrix form	18
4.3	The GLS fit	19
5	The simplest models	19
5.1	One-way ANOVA with random effects	19
5.2	Estimating the one-way ANOVA model	21
5.2.1	Mixed model approach	24
5.3	EBLUPs	25
6	Slightly more complex models	27
6.1	Means as outcomes regression	27
6.2	One-way ANCOVA with random effects	28
6.3	Random coefficients model	28
6.4	Intercepts and Slopes as outcomes	29
6.5	Nonrandom slopes	30
6.6	Asking questions: CONTRAST and ESTIMATE statement	30
7	A second look at multilevel models	34
7.1	What is a mixed model really estimating	34
7.2	Var($Y X$) and T	35
7.2.1	Random slope model	35
7.2.2	Two random predictors	36
7.2.3	Interpreting Chol(T)	37
7.2.4	Recentering and balancing the model	39
7.2.5	Random slopes and variance components parametrization	39
7.2.6	Testing hypotheses about T	40

7.3	Examples	43
7.4	Fitting a multilevel model: contextual effects	44
7.4.1	Example	45
7.5	Diagnostics	53
8	Longitudinal Data	55
8.1	The basic model	56
8.2	Analyzing longitudinal data	63
8.2.1	Classical or Mixed models	63
8.3	Pothoff and Roy	65
8.3.1	Univariate ANOVA	68
8.3.2	MANOVA repeated measures	70
8.3.3	Random Intercept Model with Autocorrelation	73
8.3.4	Comparing Different Covariance Models	76
8.3.5	Exercises on Pothoff and Roy	76
8.4	Panel Study of Income Dynamics (PSID)	77
8.5	Non-linear Growth Curves with NLMIXED	81
8.6	Logistic Mixed Regression with NLMIXED	81
9	Bibliography	82
A	Synopsis of SAS commands in PROC MIXED	85
A.1	PROC MIXED all-dressed	87
A.2	Exercises	88
B	TODO	91
C	A synopsis of topics for day 2:	96
D	Changes	96

1 Introduction

The last decade has seen rapid growth in the development and use of models suitable for multilevel or longitudinal data. We can identify at least four broad approaches: the use of derived variables, econometric models, latent trajectory models using structural equations models and mixed models. This course will focus on the use of mixed models for multilevel and longitudinal data.

They have a wide potential for applications to data that are otherwise awkward or impossible to model. Some key applications are to nested data structures, e.g. students within classes within schools within school boards, and to longitudinal data, especially ‘messy’ data where measurements are taken at irregular times, e.g. clinical data from patients measured at irregular times or panel data with changing membership. Another application is to panel data with time-varying covariates, e.g. age where each cohort has a variety of ages and the researcher is interested in studying age effects.

New books and articles on multilevel and longitudinal models are being released much faster than one can read or afford them. One way to get started for someone who intends to start with SAS would be to read:

- Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling Snijders and Bosker (1999) which is an excellent up-to-date book with a broad conceptual coverage in conjunction with
- “Using SAS PROC MIXED to fit Multilevel Models, Hierarchical Models, and Individual Growth Models” by Judith Singer Singer (1998).
- Books like Littell et al. *SAS System for Mixed Models* Littell et al. (1996) and Verbeke and Molenberghs (1997) would be good sequels.
- If you want to deepen the basic concepts, there is an excellent introductory book by Kreft and de Leeuw although it refers to HLM instead of SAS: Kreft and de Leeuw (1998)

Other books recently published are also worthy of note:

- Applied Mixed Models in Medicine: Statistics in Practice Helen Brown (2000).
- Generalized, Linear, and Mixed Models McCulloch and Searle (2000). [illustrious author but recent scathing review]
- Longitudinal Data Analysis : Designs, Models and Methods Bijleveld and Van der Kamp (1998). [Inexpensive paperback edition]
- Mixed Effects Models in S and S-Plus Pinheiro and Bates (2000).

Two classics:

- Bryk and Raudenbush (1992)
- Multilevel Statistical Models by Harvey Goldstein now available on-line (see below).

The choice of software to fit mixed models continues to grow but there are a number of well established niches. It is safe to say that no package is uniformly superior to all the others. Each package has strengths and weaknesses, otherwise we wouldn't have to mention more than one.

- An excellent special-purpose programme is MLwiN developed by a group associated with Harvey Goldstein at the Institute of Education at the University of London. It uses a graphical interface that is very appealing for multilevel modelling and produces some very interesting plots quite easily. Its basic linear methods are well integrated with more advanced methods: logistic mixed models, MCMC estimation, to make the transition relatively easy. One shortcoming is its unstructured parametrization of the variance matrix of random effects. MLwiN seems to be able to handle very large data sets as long as they fit in RAM.
- NLME by Bates and Pinheiro is a library in S-Plus. This is my favourite working environment but it's best for those who enjoy programming. It's a good environment for non-linear normal error models and can be adapted to fit logistic hierarchical models. S-Plus is very strong for graphics and for its programmability which allows you to easily refine and reuse analytic solutions.
- PROC MIXED in SAS is a solid workhorse. The ability to specify a wide variety of structures for both the within-cluster covariance matrix and for the random effects covariance matrix is an important feature. NLMIXED is a recent and interesting addition. SAS is still very weak (in my opinion) in graphics but this is being remedied to a large extent by the work of my colleague Michael Friendly. It was selected for this programme because of numerous requests for courses in mixed models using SAS and because of its wide availability in academic and governmental circles.
- GLLAMM in STATA is a very interesting programme in development. It might not be well suited to very large datasets but one should check.
- HLM developed by Bryk and Raudenbush is popular among researchers in education. I don't find its syntax inspiring but it has many devoted users.

Some good resources on the web include:

- Two multilevel statistics books are available free online: The Internet Edition of Harvey Goldstein's extensive but challenging *Multilevel Statistical Models* can be downloaded from <http://www.arnoldpublishers.com/support/goldstein.htm>

- *Applied Multilevel Analysis* by Joop Hox is available at <http://www.fss.uu.nl/ms/jh/publist/amaboek.pdf>.
- Multilevel Modelling Newsletters produced twice yearly by The Multilevel Models Project in the Institute of Education at the University of London: <http://multilevel.ioe.ac.uk/publref/newsletters.html>
- The web site for the Multilevel Models Project: <http://www.ioe.ac.uk/multilevel/>
- One can subscribe to an email discussion list at <http://www.mailbase.ac.uk/lists/multilevel/>

2 Preliminaries

We begin by reviewing a few concepts from regression and multivariate data analysis. Our emphasis is on concepts that are frequently omitted, or, at least, not well explored in typical courses.

2.1 Is β_1 the effect of X_1 ?

Consider the familiar regression equation:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad (1)$$

To make this more concrete, suppose we are studying the relationship between $Y = Health$, $X_1 = Weight$ and $X_2 = Height$.

$$E(Health) = \beta_0 + \beta_1 Weight + \beta_2 Height \quad (2)$$

so that β_1 is the ‘effect’ of changing *Weight*. What if we are really interested in the ‘effect’ of *ExcessWeight*. Maybe we should replace *Weight* with *ExcessWeight*. Let’s suppose that

$$ExcessWeight = Weight - (\phi_0 + \phi_1 Height) \quad (3)$$

where $\phi_0 + \phi_1 Height$ is the ‘normal’ *Weight* for a given *Height*. What happens if we fit the model:

$$E(Health) = \gamma_0 + \gamma_1 ExcessWeight + \gamma_2 Height \quad (4)$$

where we now expect γ_1 to be the effect of *ExcessWeight*.

If we substitute the definition of *ExcessWeight* in 3 into 4, we get:

$$E(\textit{Health}) = \gamma_0 + \gamma_1 \textit{ExcessWeight} + \gamma_2 \textit{Height} \quad (5)$$

$$= \gamma_0 + \gamma_1 (\textit{Weight} - (\phi_0 + \phi_1 \textit{Height})) + \gamma_2 \textit{Height} \quad (6)$$

$$= (\gamma_0 - \gamma_1 \phi_0) + \gamma_1 \textit{Weight} + (\gamma_2 - \gamma_1 \phi_1) \textit{Height} \quad (7)$$

and, using the fact that identical linear functions have identical coefficients:

$$\gamma_0 = \beta_0 + \beta_1 \phi_0 \quad (8)$$

$$\gamma_1 = \beta_1 \quad (9)$$

$$\gamma_2 = \beta_2 + \beta_1 \phi_1 \quad (10)$$

The coefficient we expected to change is the only one that stays the same!

This illustrates the importance of thinking of β_1 in 2 as the effect of changing *Weight* **keeping *Height* constant**, which turns out to be the same thing as changing *ExcessWeight* keeping *Height* constant as long as *ExcessWeight* is defined as a linear function of *Weight* and *Height*. In multiple regression, the other variables in the model are, in a sense, more important in defining the meaning of a coefficient than the variable to which the coefficient is attached, whose role may be seen at that of providing a measuring stick for units. This fact will play an important role when we consider controlling for ‘contextual variables.’ See the appendix for an explanation of this phenomenon using the matrix formulation for regression.

2.2 Visualizing multivariate variance

Understanding multivariate variance is important to unravel some mysteries of random coefficient models. With univariate random variables, it’s generally easier to think in terms of the standard deviation ($SD = \sqrt{VAR}$). What is the equivalent of the *SD* for a bivariate random vector? The easiest way of thinking about this uses the ‘concentration ellipse’ for the bivariate normal. A few of the ideas we can illustrate:

- The centre of the ‘standard ellipse’ is at the point of means.
- The shadow of the ellipse onto either axis gives mean ± 1 SD.
- The shadow of the ellipse onto ANY axis gives mean ± 1 SD for a variable represented by projection onto the axis, i.e. any linear combination of X_1 and X_2 .
- The vertical slice through the centre of the ellipse gives the residual SD of X_2 given X_1 . When X_1 and X_2 have a multivariate normal distribution, this is also the conditional SD of X_2 given X_1 .

- Mutatis mutandis for X_1 given X_2 .
- The line through the points of vertical tangency is the regression line of X_2 on X_1 .
- Mutatis mutandis for X_1 on X_2 .
- The fact that the regression line goes through the points of vertical tangency and NOT the corners of the rectangle containing the ellipse (NOR the principal axis of the ellipse) is the essence of the **regression paradox**. Indeed, the very word **regression** comes from the notion of ‘regression to mediocrity’ embodied in the fact that the regression line (i.e. the expected value of X_2 given X_1) is flatter than the ‘SD line’ (the line through the corners of the rectangle).
- The bivariate SD: Take any pair of **conjugate axes** on the ellipse as shown in the diagram below. (In 2D, axes are conjugate if each axis is parallel to the tangent of the ellipse at the point where the other axis intersects the ellipse ... a picture is worth 26 words.) Put them as column vectors in a 2×2 matrix and you have the ‘square root’ of the variance matrix, i.e. the SD for the bivariate random vector. The choice of conjugate axes is not unique – neither is the ‘square root’ of a matrix. Depending on your choice of conjugate axes, you can get various factorizations of the variance matrix: upper or lower Choleski, symmetric non-negative definite ‘square root’, principal components factorization (singular value decomposition), etc.

2.2.1 Geometry of regression

See handout

2.3 Combining estimators

2.3.1 Univariate estimators

See handout

2.3.2 Multivariate estimators

See handout

2.4 Simpson and Robinson

We first introduce the data set we will use to illustrate concepts in multilevel modelling. We use an example from Bryk and Raudenbush (1992) in which mixed effects models are presented as hierarchical linear models. This seems to me to be the easiest way to present and develop the ideas. We will use a subset of the data used in their book. It consists of data on math achievement, SES, minority status and sex of students in 40 schools. The schools are classified as “public” or “catholic”. One goal of the analysis is to study the relationship between math achievement and SES in different settings. **See figures A and B.**

Let just consider the relationship between math achievement, Y , and SES, X . Consider three ways (we’ll see more later) of estimating the relationship between X and Y .

1. An **ecological regression** of the individual school means of Y on the individual school means of X .
2. Pooled regression of Y on X ignoring school entirely. This estimates what is sometimes called the **marginal relationship** between Y and X .
3. A regression of Y on X adjusting for school, i.e. including school as a categorical variable or, equivalently, first ‘partialling out’ the effect of schools. This estimates the **conditional relationship** between Y and X .

Robinson’s paradox to the fact that (1) and (3) can have opposite signs. Simpson’s paradox refers to the fact that (2) and (3) can also have opposite signs. In fact, estimate

(2) will lie between (1) and (3). How much closer it is to (1) than to (3) depends on the variance of X and Y within schools. The following (not very realistic) diagram illustrates these relationships.

2.5 Matrix formulation of regression

We include this for completeness: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is such a universal and convenient shorthand that we need to spell out what it means for those for whom it is murky.

Consider the equation for a single observation (we'll assume 2 X variables):

$$Y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i \quad i = 1, \dots, N \quad (11)$$

with ε_i iid $N(0, \sigma^2)$. We pile these equations one on top of the other:

$$\begin{aligned} Y_1 &= \beta_0 + x_{11}\beta_1 + x_{21}\beta_2 + \varepsilon_1 \\ Y_2 &= \beta_0 + x_{12}\beta_1 + x_{22}\beta_2 + \varepsilon_2 \\ &\vdots \\ Y_i &= \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i \\ &\vdots \\ Y_N &= \beta_0 + x_{1N}\beta_1 + x_{2N}\beta_2 + \varepsilon_N \end{aligned}$$

Note that the β s remain the same from line to line but Y s X s and ε s change. Using vectors

and matrices and exploiting the rules for multiplying matrices:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1N} & x_{2N} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix} \quad (12)$$

or, in short-hand:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (13)$$

In multilevel models with, say J schools indexed by $j = 1, \dots, J$ and with the j th school having n_j students, we block together students of the same school so we can write, for the j th school:

$$\begin{bmatrix} Y_{1j} \\ Y_{2j} \\ \vdots \\ Y_{n_1j} \end{bmatrix} = \begin{bmatrix} 1 & x_{11j} & x_{21j} \\ 1 & x_{12j} & x_{22j} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n_1j} & x_{2n_1j} \end{bmatrix} \begin{bmatrix} \beta_{0j} \\ \beta_{1j} \\ \beta_{2j} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1j} \\ \varepsilon_{2j} \\ \vdots \\ \varepsilon_{n_1j} \end{bmatrix} \quad (14)$$

or, in short hand:

$$\mathbf{Y}_j = \mathbf{X}_j\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j \quad (15)$$

We can stack schools on top of each other. If all schools are assumed to have the same value for $\boldsymbol{\beta}_j \equiv \boldsymbol{\beta}$, then we can stack the \mathbf{X} s vertically:

$$\begin{bmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_j \\ \vdots \\ \mathbf{Y}_J \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_j \\ \vdots \\ \mathbf{X}_J \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_j \\ \vdots \\ \boldsymbol{\varepsilon}_J \end{bmatrix} \quad (16)$$

or, in shorter form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (17)$$

If the $\boldsymbol{\beta}_j$ s are different we need to stack the \mathbf{X}_j s diagonally:

$$\begin{bmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_j \\ \vdots \\ \mathbf{Y}_J \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{X}_j & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{X}_J \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_j \\ \vdots \\ \boldsymbol{\beta}_J \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_j \\ \vdots \\ \boldsymbol{\varepsilon}_J \end{bmatrix} \quad (18)$$

or, in shorter form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (19)$$

Now, you know why you see $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ so often!

Something that gets used over and over again is the fact that, if $\boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ then the OLS (ordinary least-squares) estimator is

$$\hat{\boldsymbol{\beta}}^{OLS} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y} \quad (20)$$

with variance $\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$

If the components of $\boldsymbol{\varepsilon}$ are not iid but $\boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma})$ then the GLS (generalized least-squares) estimator is

$$\hat{\boldsymbol{\beta}}^{GLS} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1} \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y} \quad (21)$$

with variance $(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}$.

3 The Multilevel Model

We develop the ideas mixed and multilevel modelling in two stages:

1. Multilevel models as presented in (Bryk and Raudenbush) in which the unobserved parameters at the lower level are modelled at the higher level. This is the representation used in *HLM*, the software developed by Bryk and Raudenbush and, to a limited extent in MLwiN.
2. Mixed model in which the levels are combined into two equations, one for ‘fixed effects’ and the other for ‘random effects.’ This is the form used in *SAS* and in most other packages.

Although the former is more complex, it seems much more natural and provides a more intuitive approach to the models.

We will use the high school Math Achievement data mentioned above as an extensive example. We think of our data as structured in two levels: *students within schools* and *between schools*.

3.1 Within school model

First suppose we are dealing with only one school. Let Y_i and X_i be the math achievement score and SES respectively of the i th student. We can formulate a model:

$$Y_i = \beta_0 + \beta_1 X_i + r_i \tag{22}$$

where β_1 is the average change in math achievement for a unit change in SES, β_0 is the

expected math achievement at $\text{SES} = 0$, and r_i is the random deviation of the i th student from the expected (linear) pattern. We assume r_i to be i.i.d. $N(0, \sigma^2)$.

3.2 Model to compare two schools

We suppose that the relationship between math achievement and SES might be different in two schools. We can index β s by school:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij} \tag{23}$$

where $j = 1, 2$ and $i = 1, \dots, n_j$ where n_j is the number of students measured in school j . Again, we assume r_{ij} to be i.i.d. $N(0, \sigma^2)$.

We can fit this model and ask various questions:

1. Do the two schools have the same structure, i.e. are the β s the same?
2. Perhaps the intercepts are different with one school “better” than the other but the slopes are the same.
3. Perhaps the slopes are different but one school is better than the other over the range of X .
4. Maybe there’s an essential interaction with one school better for high X and the other for low X .
5. We could also allow different σ s and test equality of variance (conditional given X).
6. We should also be ready to question the linearity implied by the model.

3.3 Within school model for many schools

This is similar to two schools but $j = 1, \dots, J$.

3.4 Between school model

This is also known as a “Level 2” model. We think of each school as characterized by its own pair of values for β_0 and β_1 which are, in turn, modeled as depending on “between school” (or “outer” or “Level 2”) variables.

This table of synonyms is taken from Snijders and Bosker (1999):

macro-level units	micro-level units
macro-units	micro-units
primary units	secondary units
clusters	elementary units
level-2 units	level-1 units

Variables can be classified as **macro variables** that describe the macro-units and have the same value for each micro-unit within a macro unit and **micro variables** that do vary between micro-units within macro-units. Some synonyms:

macro variable	micro variable
macro-level predictor	micro-level variable
outer variable	inner variable
level-2 variable	level-1 variable

Let W represent an “outer” variable. In this example we use an indicator for Catholic schools: W_j is equal to 1 if school j is Catholic and 0 if it is public.

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j} \quad (24)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j} \quad (25)$$

where:

1. γ_{00} is the mean achievement intercept for public schools.
2. γ_{01} is the mean difference in intercept between catholic and public schools.
3. γ_{10} is the mean slope in public schools.
4. γ_{11} is the mean difference in (or difference in mean) slope between catholic and public schools.
5. u_{0j} is the unique “effect” of school j on the achievement intercept, conditional given W .

6. u_{1j} is the unique “effect” of school j on the slope, conditional given W .

Now, u_{0j} and u_{1j} are Level 2 random variables (random effects) which we assume to have 0 mean and variance-covariance matrix:

$$T = \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix}$$

This is a multivariate model with the complication that the dependent variables, β_{0j} , β_{1j} are not directly observable. One way to proceed would be to use a two-stage process:

1. Estimate β_{0j} , β_{1j} with least-squares within each school, and
2. use the estimated values in a Level-2 analysis with the model above.

Some problems with this approach are:

1. Each $\hat{\beta}_{0j}$, $\hat{\beta}_{1j}$ might have a different estimation variance due to differing n_j s and differing predictor matrices X_j in each school. A Level 2 analysis that uses OLS will not take these factors in consideration.
2. Even if X_j (thus n_j) is the same for each school, we might be interested in getting information on \mathbf{T} , not

$$\text{Var}(\hat{\boldsymbol{\beta}}_j) = \mathbf{T} + \sigma^2(\mathbf{X}'_j\mathbf{X}_j)^{-1} \tag{26}$$

3. $\hat{\beta}_{0j}$ and $\hat{\beta}_{1j}$ might be reasonable estimates of the *parameters* β_0 and β_1 but, as ‘estimators’ of the random variables β_0 and β_1 they ignore the information contained in the distribution of β_0 and β_1 .
4. Some Level 1 models might not be estimable, so information from these units is lost.

4 Combined model

4.1 Multilevel form

We combine the models by substituting the *between school* model into the *within school* model:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij} \quad (27)$$

$$= \gamma_{00} + \gamma_{01}W_j + u_{0j} \quad (28)$$

$$+ (\gamma_{10} + \gamma_{11}W_j + u_{1j})X_{ij} + r_{ij} \quad (29)$$

$$= \gamma_{00} + \gamma_{01}W_j + \gamma_{10}X_{ij} + \gamma_{11}W_jX_{ij} \quad (30)$$

$$+ u_{0j} + u_{1j}X_{ij} + r_{ij} \quad (31)$$

This looks like the sum of two linear models:

$$\gamma_{00} + \gamma_{01}W_j + \gamma_{10}X_{ij} + \gamma_{11}W_jX_{ij} \quad (32)$$

with parameters $\gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11}$, and:

$$u_{0j} + u_{1j}X_{ij} \quad (33)$$

with random ‘parameters’ u_{0j} and u_{1j} .

Another way of looking at this model is to see it as a linear model with a complex form of error. Letting

$$\delta_{ij} = u_{0j} + u_{1j}X_{ij} + r_{ij} \quad (34)$$

we can write the model as:

$$Y_{ij} = \gamma_{00} + \gamma_{01}W_j + \gamma_{10}X_{ij} + \gamma_{11}W_jX_{ij} + \delta_{ij} \quad (35)$$

where δ_{ij} s are **not** identically $N(0, \sigma^2)$ and are **not** independent since the same u_{0j} and u_{1j} contribute to random error for all δ_{ij} s in the j th school.

If T and σ^2 were known then the variance-covariance matrix of the random error could be computed and the model fitted with Generalized Least-Squares (GLS).

With T and σ^2 unknown, we can iteratively estimate them and use the estimated values to fit the linear parameters, γ_{st} by GLS. There are variants depending on the way in which T and σ^2 are estimated. Using full likelihood yields what is often called ‘‘IGLS,’’ ‘‘ML,’’ or ‘‘FIML.’’ Using the conditional likelihood of residuals given \hat{Y} yields ‘‘RIGLS’’ or ‘‘REML’’ (R for restricted, reduced or ...).

4.2 Matrix form

Take all observations in school j and assemble them into vectors and matrices: (this is called the Laird-Ware formulation of the model from Laird and Ware (1982))

$$\mathbf{Y}_j = \mathbf{X}_j\boldsymbol{\gamma} + \mathbf{Z}_j\mathbf{u}_j + \mathbf{r}_j \quad (36)$$

where

$$\mathbf{Y}_j = \begin{bmatrix} Y_{1j} \\ \vdots \\ Y_{n_jj} \end{bmatrix} \quad \mathbf{X}_j = \begin{bmatrix} 1 & W_j & X_{1j} & W_j X_{1j} \\ 1 & W_j & X_{2j} & W_j X_{2j} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & W_j & X_{n_jj} & W_j X_{n_jj} \end{bmatrix} \quad \mathbf{Z}_j = \begin{bmatrix} 1 & X_{1j} \\ 1 & X_{2j} \\ \vdots & \vdots \\ 1 & X_{n_jj} \end{bmatrix} \quad (37)$$

$$\mathbf{u}_j = \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \quad \boldsymbol{\gamma} = \begin{pmatrix} \gamma_{00} \\ \gamma_{01} \\ \gamma_{10} \\ \gamma_{11} \end{pmatrix} \quad \mathbf{r}_j = \begin{pmatrix} r_{1j} \\ r_{2j} \\ \vdots \\ r_{n_jj} \end{pmatrix} \quad j = 1, \dots, J \quad (38)$$

The distribution of the random elements is: $\mathbf{u}_j \sim N(0, T)$, $\mathbf{r}_j \sim N(0, \sigma^2 I)$ with \mathbf{u}_j independent of \mathbf{r}_j .

Now we put the school matrices together into big matrices:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{Z}\mathbf{u} + \mathbf{r} \quad (39)$$

where

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_J \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_J \end{bmatrix} \quad \mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_J \end{bmatrix} \quad \mathbf{r} = \begin{bmatrix} \mathbf{r}_1 \\ \vdots \\ \mathbf{r}_J \end{bmatrix} \quad (40)$$

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{Z}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{Z}_J \end{bmatrix} \quad (41)$$

with

$$\mathbf{u} \sim N(0, \begin{bmatrix} T & 0 & \cdots & 0 \\ 0 & T & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & T \end{bmatrix}) \quad (42)$$

and

$$\mathbf{r} \sim N(0, \sigma^2 I) \quad (43)$$

which might be deceptive because the “ T ” is now much larger than before. The new block diagonal matrix for the variance of \mathbf{u} is sometimes expressed with the same symbol as the variance of \mathbf{u}_j . To avoid confusion we can use \ddot{T} .

4.3 The GLS fit

With the matrix formulation of the model, it is easy to express the GLS estimator of $\boldsymbol{\gamma}$. First denote:

$$V = \text{Var}(\boldsymbol{\delta}) = \mathbf{Z}\ddot{T}\mathbf{Z}' + \sigma^2 I \quad (44)$$

Then the GLS estimator can be expressed as:

$$\hat{\boldsymbol{\gamma}} = (\mathbf{X}'V^{-1}\mathbf{X})^{-1}\mathbf{X}'V^{-1}\mathbf{Y} \quad (45)$$

We will see that the presence of V^{-1} can result in an estimate that is very different from its OLS analogue.¹

5 The simplest models

5.1 One-way ANOVA with random effects

This is the simplest random effects models and provides a good starting point to illustrate the special characteristics of these models.

¹One ironic twist concerns small estimated values of σ^2 . Normally this is a cause for rejoicing; however it can result in a nearly singular V . Although this need not imply that $\mathbf{X}'V^{-1}\mathbf{X}$ is nearly singular, some algorithms seem not to take advantage of this.

Level 1 model:

$$Y_{ij} = \beta_{0j} + r_{ij} \quad (46)$$

Level 2 model:

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad (47)$$

Combined model:

$$Y_{ij} = \gamma_{00} + u_{0j} + r_{ij} \quad (48)$$

$$\text{Var}(Y_{ij}) = \text{Var}(u_{0j} + r_{ij}) \quad (49)$$

$$= \tau_{00} + \sigma^2 \quad (50)$$

Note the intraclass correlation coefficient:

$$\rho = \tau_{00}/(\tau_{00} + \sigma^2) \quad (51)$$

Also note that within each school:

$$E(\bar{Y}_{.j}|\beta_{0j}) = \beta_{0j} \quad (52)$$

$$\text{Var}(\bar{Y}_{.j}|\beta_{0j}) = \sigma^2/n_j \quad (53)$$

but across the population:

$$E(\bar{Y}_{.j}) = \gamma_{0j} \quad (54)$$

$$\text{Var}(\bar{Y}_{.j}) = \tau_{00} + \sigma^2/n_j \quad (55)$$

This is an example of two very useful facts:

1. the unconditional (sometimes called ‘marginal’ but not by economists) mean is equal to the **mean conditional mean**,
2. the unconditional variance is equal to the **mean of the conditional variance** plus the **variance of the conditional mean**, i.e.:

$$\text{Var}(\bar{Y}_{.j}) = \text{E}(\text{Var}(\bar{Y}_{.j}|\beta_{0j})) + \text{Var}(\text{E}(\bar{Y}_{.j}|\beta_{0j})) \quad (56)$$

$$= \sigma^2/n_j + \text{Var}(\beta_{0j}) \quad (57)$$

$$= \sigma^2/n_j + \tau_{00} \quad (58)$$

$$(59)$$

$$\text{Var}(Y_{ij}) = \text{E}(\text{Var}(Y_{ij}|\beta_{0j})) + \text{Var}(\text{E}(Y_{ij}|\beta_{0j})) \quad (60)$$

$$= \sigma^2 + \text{Var}(\beta_{0j}) \quad (61)$$

$$= \sigma^2 + \tau_{00} \quad (62)$$

5.2 Estimating the one-way ANOVA model

There are three kinds of parameters that need to be estimated:

1. **fixed effect parameters:** in this case there is only one: γ_{00} ,
2. **variance-covariance components:** τ_{00} and σ^2 ,
3. **random effects:** β_{0j} or, equivalently, combined with γ_{00} : u_{0j} .

We use a different approach for each type of parameter.

The **fixed effects parameters** are like linear regression parameters except that they are estimated from observations that are not independent. Instead of using OLS (ordinary least-squares) we use **GLS (generalized least-squares)** using the estimates of the variance-covariance components as the variance matrix in the GLS procedure.

The **variance-covariance parameters** are estimated using **ML (maximum likelihood) or REML (restricted maximum likelihood)**.

Note that each step above assumes that the other one has been completed. What really happens is that estimation goes back and forth between the two steps until convergence.

The **random effects** are not just parameters. They are realizations of random variables. This means that we have two sources of information about them: we can ‘estimate’ them from the observed data and we can ‘guess’ them from their distribution. Putting these two sources of information together is the essence of Bayesian estimation, or **empirical Bayesian** estimation because the distribution of the random effects, determined by T , is estimated from the data and model. The random effects are **predicted** using **EBLUPs (Empirical Best Linear Unbiased Predictors)** with the empirical **posterior expectation**:

$$E(\beta_{01}, \dots, \beta_{0J} | Y_1, \dots, Y_n) \quad (63)$$

i.e. the expected value of what is unknown given what is known.

We will look at the estimation of the three types of parameters in detail in this example.

First we consider the analysis of the data using OLS in which we treat $\beta_{01}, \dots, \beta_{0J}$ as non-random parameters. **The coding of the school effect determines what is estimated by the intercept term.** It is a weighted linear combination of the β_{0j} s:

$$\psi_w = \sum_{j=1}^J w_j \beta_{0j} \quad (64)$$

If the coding uses “true” contrasts (each column of the **coding matrix** sums to 0) the weights are all equal to $1/J$ and ψ_w is the ordinary mean of the β_{0j} s:

$$\psi_w = \frac{1}{J} \sum_{j=1}^J \beta_{0j} \quad (65)$$

In this case

$$\hat{\psi}_w = \frac{1}{J} \sum_{j=1}^J \bar{Y}_j = \bar{Y}_{Schools} \quad (66a)$$

With “sample size” coding, e.g.

$$\begin{array}{rcccccc}
 & V_1 & V_2 & V_3 & \dots & V_{J-1} \\
 School_1 & n_J & 0 & 0 & \dots & 0 \\
 School_2 & 0 & n_J & 0 & \dots & 0 \\
 School_3 & 0 & 0 & n_J & \dots & 0 \\
 School_4 & 0 & 0 & 0 & \dots & 0 \\
 \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
 School_{J-1} & 0 & 0 & 0 & \dots & n_J \\
 School_J & -n_1 & -n_2 & -n_3 & \dots & -n_{J-1}
 \end{array} \tag{67}$$

each column of the **design matrix** sums to 0 and the intercept will estimate:

$$\psi_w = \frac{\sum_{j=1}^J n_j \beta_{0j}}{\sum_{j=1}^J n_j} \tag{68}$$

which weights each school according to its sample size. This can be thought of as the mean of the population of **students** instead of the population of **schools**. The estimator would be the overall average of Y :

$$\hat{\psi}_w = \frac{\sum_{j=1}^J n_j \bar{Y}_j}{\sum_{j=1}^J n_j} = \bar{Y}_{..} = \bar{Y}_{Students} \tag{69}$$

We are not limited to these two obvious choices. A more appropriate set of weights could be **school size**, with coding:

$$\begin{array}{rcccccc}
 & V_1 & V_2 & V_3 & \dots & V_{J-1} \\
 School_1 & s_J & 0 & 0 & \dots & 0 \\
 School_2 & 0 & s_J & 0 & \dots & 0 \\
 School_3 & 0 & 0 & s_J & \dots & 0 \\
 School_4 & 0 & 0 & 0 & \dots & 0 \\
 \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
 School_{J-1} & 0 & 0 & 0 & \dots & s_J \\
 School_J & -s_1 & -s_2 & -s_3 & \dots & -s_{J-1}
 \end{array}$$

the intercept would estimate:

$$\psi_s = \frac{\sum_{j=1}^J s_j \beta_{0j}}{\sum_{j=1}^J s_j}. \tag{70}$$

In each case the form of the estimate is a weighted mean of the individual school averages:

$$\hat{\psi}_w = \sum_{j=1}^J w_j \bar{Y}_j \quad (71)$$

with variance:

$$\text{Var}(\hat{\psi}_w | \beta_{01}, \dots, \beta_{0J}) = \sum_{j=1}^J w_j^2 \frac{\sigma^2}{n_j} \quad (72)$$

where the weights, w_j s, sum to 1. Note that the variance is minimized when the weights are proportional to n_j , i.e. $w_j = n_j/n$ where n is the total sample size: $n = \sum_j n_j$. In this case the variance is σ^2/n . Thus, the **student mean** is the parameter estimated with the least variance.

5.2.1 Mixed model approach

With a mixed model we want to estimate γ_{00} instead of a particular linear combination of β_{0j} s. Any weighted mean $\hat{\psi}_w = \sum_j w_j \bar{Y}_j$ of \bar{Y}_j s will be unbiased for γ_{00} because

$$\text{E}(\hat{\psi}_w) = \text{E}\left(\sum_j w_j \bar{Y}_j\right) \quad (73)$$

$$= \sum_j w_j \text{E}(\beta_{0j}) \quad (74)$$

$$= \sum_j w_j \gamma_{00} = \gamma_{00} \quad (75)$$

if the w_j s are weights with $\sum_j w_j = 1$.

Now, to calculate the variance of \bar{Y}_j as an estimator of γ_{00} we first need the variance of \bar{Y}_j as an estimator of γ_{00} with β_{0j} random:

$$\text{Var}(\bar{Y}_j) = \tau_{00} + \sigma^2/n_j \quad (76)$$

Thus:

$$\text{Var}(\hat{\psi}_w) = \sum_j w_j^2 (\tau_{00} + \sigma^2/n_j) \quad (77)$$

The optimal estimator is obtained by taking weights **inversely proportional** to $(\tau_{00} + \sigma^2/n_j)$.

Consider the implications: **if τ_{00} is much larger than σ^2** , the weights will be nearly constant and $\hat{\psi}_w$ will be close to $\bar{Y}_{Schools}$

Conversely, if τ_{00} is **much smaller than** σ^2 , the weights will be nearly proportional to n_j and the estimator will be close to $\bar{Y}_{Students}$.

If it is not reasonable to treat the β_{0j} s as a random sample from the same $N(0, \tau_{00})$ distribution then these two estimators could estimate two quantities with very different meanings. Consider, for example, what would happen if there is a strong relationship between β_{0j} and n_j . What gets estimated is governed by the ratio τ_{00}/σ^2 – a purely statistical consideration quite disconnected from any interpretation of the estimator. It is important to appreciate that your estimator is determined by considerations that might not be relevant.

In SAS, the (minimal) commands would be:

```
PROC MIXED;
  CLASS SCHOOL;
  MODEL Y = ;
  RANDOM INTERCEPT / SUBJECT=SCHOOL;
  RUN;
```

In MLwiN, you set x_0 equal to a “constant variable” (call it *cons*) all of whose values are 1 and set the coefficient of *cons* so it has a **fixed** component, a **level-2 random** component and a **level-1 random** component.² This produces a model of the form:

$$y_{ij} = \beta_{0ij}x_0 \tag{78}$$

$$\beta_{0ij} = u_{0j} + e_{0ij} \tag{79}$$

Note the curious way in which individual error is introduced as a random coefficient of the constant term. This can be in that the individual error could be associated with a variable predictor thereby allowing a form of weighting.

5.3 EBLUPs

Estimating the u_{0j} s involves using two sources of information: the data and their distribution as random variables. First consider the OLS estimator for β_{0j} :

$$\hat{\beta}_{0j} = \bar{Y}_{.j} \tag{80}$$

Now, to get the *Empirical Best Linear Unbiased Predictor* of u_{0j} s, we pretend that the estimated values of γ_{00} , τ_{00} and σ^2 are the “true” values and we calculate the conditional

²All variables are numerical in MLwiN. To model a categorical predictor you must generate the numerical coding variables.

expectation of u_{0j} s given y_{ij} s. This is done most easily using the matrix formulation of the model and a formula for the conditional expectation in the multivariate case. We use partitioned matrices to express the joint distribution of \mathbf{Y} and \mathbf{u} :

$$\begin{bmatrix} \mathbf{Y} \\ \mathbf{u} \end{bmatrix} \sim N \left(\begin{bmatrix} \mathbf{X}\boldsymbol{\gamma} \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{Z}\ddot{\mathbf{T}}\mathbf{Z}' + \sigma^2\mathbf{I} & \mathbf{Z}\ddot{\mathbf{T}} \\ \ddot{\mathbf{T}}\mathbf{Z}' & \ddot{\mathbf{T}} \end{bmatrix} \right)$$

A “well-known” formula gives:

$$\hat{\mathbf{E}}(\mathbf{u}|\mathbf{Y}) = \ddot{\mathbf{T}}\mathbf{Z}\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\gamma})$$

where $\mathbf{V} = \mathbf{Z}\ddot{\mathbf{T}}\mathbf{Z}' + \sigma^2\mathbf{I}$. This formula with a bit more mechanical work will give us the EBLUP below, but we will derive it intuitively:

1. We could estimate u_{0j} with the “obvious” OLS estimate:

$$\hat{u}_{0j} = \hat{\beta}_{0j} - \hat{\gamma}_{00} = \bar{Y}_{.j} - \hat{\gamma}_{00} \quad (81)$$

As an estimate of u_{0j} this has variance σ^2/n_j .

2. We could also guess that u_{0j} is equal to 0 (the mean of its distribution) and our guess would have variance τ_{00} .

How can we “best” combine these independent sources of information? By using weights proportional to inverse variance! This gives us the EBLUP of u_{0j} :

$$\tilde{u}_{0j} = \frac{\frac{1}{\sigma^2/n_j}\hat{u}_{0j} + \frac{1}{\tau_{00}}0}{\frac{1}{\sigma^2/n_j} + \frac{1}{\tau_{00}}} = \frac{\hat{u}_{0j}}{1 + \frac{\sigma^2/n_j}{\tau_{00}}} \quad (82)$$

This has the effect of **shrinking** \hat{u}_{0j} towards 0 by a factor of

$$\frac{\frac{1}{\sigma^2/n_j}}{\frac{1}{\sigma^2/n_j} + \frac{1}{\tau_{00}}} = \frac{1}{1 + \frac{\sigma^2/n_j}{\tau_{00}}} \quad (83)$$

Consider how the amount of shrinking depends on the relative values of σ^2 , τ_{00} and n_j . There will be more shrinkage if

1. τ_{00} is small: i.e. the distribution of u_{0j} is known to be close to 0.
2. σ^2 is large: i.e. $\bar{y}_{.j}$ has large variation as an estimate of β_{0j} .

3. n_j is small: ditto.

The EBLUP estimator of β_{0j} (we'll call it $\tilde{\beta}_{0j}$) works exactly the same way with the OLS estimator (analyzing each school separately) which gets shrunk towards the overall estimator $\hat{\gamma}_{00}$. This is in exactly the same spirit as shrinkage estimators derived from Bayesian, Empirical Bayes or frequentist approaches. Bradley Efron and Carl Morris wrote an interesting article on the topic in *Scientific American* Efron and Morris (1977).

6 Slightly more complex models

6.1 Means as outcomes regression

Level 1 model:

$$Y_{ij} = \beta_0 + \beta_{0j} + r_{ij} \quad (84)$$

Level 2 model:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j} \quad (85)$$

Combined model:

$$Y_{ij} = \gamma_{00} + \gamma_{01}W_j + u_{0j} + r_{ij} \quad (86)$$

Note that

$$\text{Var}(Y_{ij}) = \text{Var}(u_{0j} + r_{ij}) \quad (87)$$

as above but, in this model, $\text{Var}(Y_{ij})$ is *conditional* variance, conditional given W .

In SAS, the commands for means as outcomes model would be:

```
PROC MIXED;
  CLASS SCHOOL;
  MODEL Y = W ;
  RANDOM INTERCEPT / SUBJECT = SCHOOL;
  RUN;
```

In MLwiN, x_0 and its coefficient are set as they are for the one-way ANOVA model (see equation 78). The W variable is added as a **fixed** variable. Note that MLwiN automatically examines the new variable to determine whether it is an inner or an outer variable. The indices of the variable are set accordingly: W_{ij} for an inner variable and W_j for an outer variable. The resulting model is:

$$y_{ij} = \beta_{0ij}x_0 + \beta_1W_j \quad (88)$$

$$\beta_{0ij} = \beta_0 + u_{0j} + e_{0ij} \quad (89)$$

6.2 One-way ANCOVA with random effects

Level 1 model:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij} \quad (90)$$

Level 2 model:

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad (91)$$

$$\beta_{1j} = \gamma_{10} \quad (92)$$

Combined model:

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + u_{0j} + r_{ij} \quad (93)$$

In SAS, the commands for one-way ANCOVA with random effects are:

```
PROC MIXED;  
  CLASS SCHOOL;  
  MODEL Y = X ;  
  RANDOM INTERCEPT / SUBJECT = SCHOOL;  
  RUN;
```

In MLwiN, everything is done in the same way as for the previous model. The one exception is that MLwiN recognizes X as an inner variable and shows the index appropriately:

$$y_{ij} = \beta_{0ij}x_0 + \beta_1X_{ij} \quad (94)$$

$$\beta_{0ij} = \beta_0 + u_{0j} + e_{0ij} \quad (95)$$

6.3 Random coefficients model

Level 1 model:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij} \quad (96)$$

Level 2 model:

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad (97)$$

$$\beta_{1j} = \gamma_{10} + u_{1j} \quad (98)$$

with:

$$\text{Var} \left(\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \right) = T = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix} \quad (99)$$

Combined model:

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} \quad (100)$$

$$+ u_{0j} + u_{1j}X_{ij} + r_{ij} \quad (101)$$

In SAS, the commands for the random coefficients model are:

```
PROC MIXED;
  CLASS SCHOOL;
  MODEL Y = X ;
  RANDOM INTERCEPT X / SUBJECT = SCHOOL;
  RUN;
```

In MLwiN, the model is created similarly to the previous one except that the coefficient of X is set so that it has both a fixed and a Level-2 random component:

$$y_{ij} = \beta_{0ij}x_0 + \beta_{1j}X_{ij} \quad (102)$$

$$\beta_{0ij} = \beta_0 + u_{0j} + e_{0ij} \quad (103)$$

$$\beta_{1j} = \beta_1 + u_{1j} \quad (104)$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N \left(0, \begin{bmatrix} \sigma_{u0}^2 & \\ \sigma_{u20} & \sigma_{u1}^2 \end{bmatrix} \right) \quad (105)$$

Note that the upper-right element of the variance matrix is shown as a blank because the matrix is symmetric.

6.4 Intercepts and Slopes as outcomes

This corresponds to the full model presented above in (27).

The SAS commands for this model are:

```
PROC MIXED;
  CLASS SCHOOL;
  MODEL Y = X W X*W;
  RANDOM INTERCEPT X / SUBJECT = SCHOOL;
  RUN;
```

Note the $X*W$ term. It is called a *cross-level interaction*. It has the function of allowing the mean slope with respect to X to vary with W .

In MLwiN, we need to create an “interaction” variable that is the product of X and W , which we will call XxW . The resulting model is:

$$y_{ij} = \beta_{0ij}x_0 + \beta_1W_j + \beta_{2j}X_{ij} + \beta_3XxW_{ij} \quad (106)$$

$$\beta_{0ij} = \beta_0 + u_{0j} + e_{0ij} \quad (107)$$

$$\beta_{2j} = \beta_2 + u_{2j} \quad (108)$$

6.5 Nonrandom slopes

Consider the full model but with $\tau_{11} = 0$ (hence $\tau_{01} = 0$ also, otherwise T would not be a variance matrix). This is a model in which the variation in $\hat{\beta}_{1j}$ from school to school is wholly consistent with the expected variation *within schools* and there is no need to postulate that $\tau_{11} > 0$.

It is left as an exercise to specify the SAS and MLwiN commands to produce this model:

6.6 Asking questions: CONTRAST and ESTIMATE statement

The CONTRAST and ESTIMATE statements in SAS allow you to ask specific questions about the model. We will use the ‘Intercept and slopes as outcome’ model above and consider asking some specific questions such as

Is there a significant difference between Public and Catholic schools at various values of SES, e.g. some meaningful quantiles

Proportion	10%	50%	90%
Quantile of SES	-0.889	0.192	1.162

There are a few steps that make it easier to ask a clear question:

1. Sketch the fitted model

2. Identify what you want to ask on the sketch.

3. Transform your question into coefficients for terms in the model.

4. Write a suitable CONTRAST or ESTIMATE statement.

We will first apply this approach to the question: Is there a significant difference between the two sectors at the 10%-quantile.

Let's have a look at a summary of the data:

	SCHOOL	MINORITY	FEMALE	SES	MATHACH	SIZE	SECTOR	MEANSES
Min.	1317	0.0000	0.000	-2.3280	-2.832	153	0.0000	-0.7510
1st Qu.	2995	0.0000	0.000	-0.4180	7.782	493	0.0000	-0.1010
Median	5783	0.0000	1.000	0.1920	13.620	1068	0.0000	0.1985

Mean	5453	0.1779	0.561	0.1481	13.110	1070	0.4203	0.1541
3rd Qu.	8202	0.0000	1.000	0.7845	18.700	1523	1.0000	0.4480
Max.	9586	1.0000	1.000	1.8320	24.990	2403	1.0000	0.7590

Then we fit the ‘intercept and slopes as outcomes’ with random slopes model and have a look at estimated coefficients. We replace Y and X with the names of the variables in the data set.

```
PROC MIXED DATA = SPIDA.HS;
  CLASS SCHOOL;
  MODEL MATHACH = SES SECTOR SES*SECTOR / S DDFM=SATTERTH;
  RANDOM INTERCEPT SES / SUBJECT = SCHOOL TYPE = FAO(2);
  RUN;
```

Note that we use the ‘S’ option in the model statement to force printing the estimated coefficients. Some of the output produced follows:

Solution for Fixed Effects					
		Standard			
Effect	Estimate	Error	DF	t Value	Pr > t
Intercept	11.6997	0.4282	31.9	27.32	<.0001
SES	2.8919	0.2659	44.7	10.88	<.0001
SECTOR	2.4715	0.7033	31.2	3.51	0.0014
SES*SECTOR	-1.1004	0.4492	14.8	-2.45	0.0273

Type 3 Tests of Fixed Effects				
	Num	Den		
Effect	DF	DF	F Value	Pr > F
SES	1	44.7	118.30	<.0001
SECTOR	1	31.2	12.35	0.0014
SES*SECTOR	1	14.8	6.00	0.0273

We specify the values of the X matrix for the two levels we want to compare:

	Intercept	SES	SECTOR	SES*SECTOR
Cath at SES = -0.889	1	-0.889	1	-0.889
Public at SES = -0.889	1	-0.889	0	0
Difference	0	0	1	-0.889

We can now modify PROC MIXED to get the desired estimate:

```
PROC MIXED DATA = SPIDA.HS;
  CLASS SCHOOL;
  MODEL MATHACH = SES SECTOR SES*SECTOR / S DDFM=SATTERTH;
  RANDOM INTERCEPT SES / SUBJECT = SCHOOL TYPE = FA0(2);
  ESTIMATE 'Sector diff at 10-pctile of SES'
    SECTOR 1 SES*SECTOR -.889 / CL E;
RUN;
```

with partial output:

Label	Estimate	Standard Error	DF	t Value	Pr > t
Sector diff at 10-pctile of SES	3.4497	0.8596	40.7	4.01	0.0003

Label	Estimates	Alpha	Lower	Upper
Sector diff at 10-pctile of SES	0.05	1.7132	5.1862	

Thus we see that the estimated MATHACH is 3.4497 higher in the Catholic sector than in the Public sector at the 10th percentile of SES.

Exercise 1 *Estimate the difference at the 90th percentile of SES.*

We now ask a question that requires 2 constraints: Are the two lines identical? The significant values for the SES*SECTOR interaction and for the SECTOR main effect suggest that they are not! However, these two tests don't constitute a formal test of the equality of the two lines. We could test equality at two points, say, SES = 1 and -1:

	Intercept	SES	SECTOR	SES*SECTOR
Cath at SES = 1	1	1	1	1
Public at SES = 1	1	1	0	0
Difference	0	0	1	1
Cath at SES = -1	1	-1	1	-1
Public at SES = -1	1	-1	0	0
Difference	0	0	1	-1

To test two constraints simultaneously, we use the CONTRAST statement and separate each constraint with a comma:

```
PROC MIXED DATA = SPIDA.HS;
  CLASS SCHOOL;
  MODEL MATHACH = SES SECTOR SES*SECTOR / S DDFM=SATTERTH;
  RANDOM INTERCEPT SES / SUBJECT = SCHOOL TYPE = FA0(2);
  CONTRAST 'No sector effect'
    SECTOR 1 SES*SECTOR 1,
    SECTOR 1 SES*SECTOR -1 / E;
RUN;
```

Note that the CONTRAST statement does not take the CL option because it only performs a test, it does not produce confidence intervals.

Exercise 2 *See if you get the same answer with different equivalent constraints: e.g. if the coefficients of SECTOR and SES*SECTOR are both 0 then the two lines are identical. Try the statement:*

```
CONTRAST 'equivalent?' SECTOR 0 , SES*SECTOR 0 / E;
```

7 A second look at multilevel models

On the second day of the course we will look more deeply into multilevel models to see what our estimators are really estimating. We will learn about the potential bias of estimators when between cluster and within cluster effects are different and we will see the role of contextual variables (e.g. cluster averages of inner variables) and within cluster residuals (i.e. inner variables centered by cluster means).

7.1 What is a mixed model really estimating

See handout.

7.2 Var($Y|X$) and \mathbf{T}

There are two ways of visualizing \mathbf{T} : directly as $\text{Var}(\mathbf{u}_j)$ or through its effect on the shape of the variance of Y as a function of the predictor X . This relationship is of more than mathematical curiosity. It is very important for an understanding of the interpretation of the components of \mathbf{T} and to understanding the meaning of hypotheses involving \mathbf{T} . We will consider the case of a single predictor and the case of two predictors.

7.2.1 Random slope model

It is easy to visualize how a random sample of lines would produce a graph with larger variance as we move to extreme values of X .

[INSERT ellipse and line graph – show correspondence]

The model is:

$$Y_{ij} = \gamma_0 + \gamma_1 X_{ij} + u_{0j} + u_{1j} X_{ij} + \varepsilon_{ij} \quad (109)$$

with $\text{Var}\left(\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix}\right) = \mathbf{T} = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix}$ and $\text{Var}(\varepsilon_{ij}) = \sigma^2$. Thus

$$\text{Var}(Y|X) = \begin{bmatrix} 1 & X \end{bmatrix} \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix} \begin{bmatrix} 1 \\ X \end{bmatrix} + \sigma^2 \quad (110)$$

$$= \tau_{00} + \sigma^2 + 2\tau_{01}X + \tau_{11}X^2 \quad (111)$$

which is a quadratic function in X . Quadratic functions have straightforward properties that we can exploit to understand the meaning of \mathbf{T} . Using calculus we can show that the minimum occurs at $X = -\tau_{01}/\tau_{11}$. The minimum is $\tau_{00} - \tau_{01}^2/\tau_{11} + \sigma^2$. An alternative way of showing this involves the purely algebraic technique of ‘completing the square’:

$$\text{Var}(Y|X) = \tau_{00} + \sigma^2 + 2\tau_{01}X + \tau_{11}X^2 \tag{112}$$

$$= (\tau_{00} - \tau_{01}^2/\tau_{11}) + \sigma^2 + \tau_{11}(X - (-\tau_{01}/\tau_{11}))^2 \tag{113}$$

This shows that forcing $\tau_{01} = 0$ is equivalent to forcing the model to have minimum variance over the origin of X , which is in general a quite arbitrary assumption. This is an assumption that violates a **principle of invariance**: the fitted model should not change when something arbitrary is changed. A familiar example of the principle of invariance is the **principle of marginality** that says that you shouldn’t drop main effects that are included in interactions in the model. The interpretation of these main effects depend on the choice of origin or coding of the other interacting variables. Of course, if 0 is not an arbitrary origin then the principle of invariance is not violated by considering whether $\tau_{01} = 0$.

7.2.2 Two random predictors

With two random predictors we can visualize the ‘iso-variance’ contours in predictor space.

$$\text{Var}(Y|X) = \begin{bmatrix} 1 & X_1 & X_2 \end{bmatrix} \begin{bmatrix} \tau_{00} & \tau_{01} & \tau_{02} \\ \tau_{10} & \tau_{11} & \tau_{12} \\ \tau_{20} & \tau_{21} & \tau_{22} \end{bmatrix} \begin{bmatrix} 1 \\ X_1 \\ X_2 \end{bmatrix} + \sigma^2 \tag{114}$$

We can show that the minimum occurs at:

$$\begin{bmatrix} x_{1 \min} \\ x_{2 \min} \end{bmatrix} = - \begin{bmatrix} \tau_{11} & \tau_{12} \\ \tau_{21} & \tau_{22} \end{bmatrix}^{-1} \begin{bmatrix} \tau_{10} \\ \tau_{20} \end{bmatrix} \quad (115)$$

and the contour ellipses have the same shape as the variance ellipses for a distribution with variance

$$\begin{bmatrix} \tau_{11} & \tau_{12} \\ \tau_{21} & \tau_{22} \end{bmatrix}^{-1} \quad (116)$$

Note then that $\tau_{12} = 0$ is equivalent to a requirement that the contour ellipse is not tilted with respect to the X_1 and X_2 axes. This could be a valid hypothesis if the variables X_1 and X_2 have a distinct meaning so that the independence of β_1 and β_2 is a plausible hypothesis. Setting τ_{12} to 0 affords an opportunity to reduce the number of parameters in \mathbf{T} . With even larger random models it becomes interesting to consider \mathbf{T} matrices of the form

$$\mathbf{T} = \begin{bmatrix} \tau_{00} & \tau_{01} & \tau_{02} & \tau_{03} \\ \tau_{10} & \tau_{11} & 0 & 0 \\ \tau_{20} & 0 & \tau_{22} & 0 \\ \tau_{30} & 0 & 0 & \tau_{33} \end{bmatrix} \quad (117)$$

These models are easy to specify in MIWin. We will later see a device with which they can be specified in SAS.

7.2.3 Interpreting Chol(\mathbf{T})

A very useful option for the RANDOM statement in PROC MIXED is ‘GC’ to print the lower triangular Choleski root of the \mathbf{T} (G in SAS) matrix. For a model with a single random

slope we have

$$\mathbf{T} = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix} \quad (118)$$

where $\sqrt{\tau_{00}}$ is the SD of the height of true regression lines when $X = 0$, $\sqrt{\tau_{11}}$ is the SD of slopes and $-\tau_{01}/\tau_{11}$ is the value of X where regression lines have minimum SD.

The lower triangular Choleski root of \mathbf{T} is a lower triangular matrix, \mathbf{C} with non-negative diagonal elements that is a ‘square root’ of \mathbf{T} in the sense that $\mathbf{C}\mathbf{C}' = \mathbf{T}$. The columns of \mathbf{C} are ‘conjugate axes’ of the variance ellipse:

The elements of the matrix are:

$$\mathbf{C} = \begin{bmatrix} \sqrt{\tau_{00}} & 0 \\ \tau_{01}/\sqrt{\tau_{00}} & \sqrt{\tau_{11} - \tau_{01}^2/\tau_{11}^2} \end{bmatrix} \quad (119)$$

whose elements may be interesting but not highly interpretable.

If we fit the model with the intercept last, however, we get more interesting results. If we specify

RANDOM SES INT / GC ... etc.

the \mathbf{T} matrix will be

$$\mathbf{T} = \begin{bmatrix} \tau_{11} & \tau_{10} \\ \tau_{01} & \tau_{00} \end{bmatrix} \quad (120)$$

and the lower-triangular Choleski root:

$$\mathbf{C} = \begin{bmatrix} c_{11} & 0 \\ c_{01} & c_{00} \end{bmatrix} = \begin{bmatrix} \sqrt{\tau_{11}} & 0 \\ \tau_{01}/\sqrt{\tau_{11}} & \sqrt{\tau_{00} - \tau_{01}^2/\tau_{11}} \end{bmatrix} \quad (121)$$

is a much more interesting matrix because its diagonal elements interpretations that are invariant with respect to relocating X :

- $\sqrt{\tau_{11}}$ is the SD of slopes
- $\sqrt{\tau_{00} - \tau_{01}^2/\tau_{11}}$ is the SD of the height of true regression lines at the value of X where this SD is minimized
- Less directly interpretable but useful: $-c_{01}/c_{11} = -\tau_{01}/\tau_{11}$ is the value of X where the minimum occurs

7.2.4 Recentering and balancing the model

If the value of X (or values of X_1 and X_2 if there are two random slopes) where the minimum SD of true regression occurs is quite far from the origin, we have seen that this induces a strong correlation in \mathbf{T} . If this is large enough, the matrix may be close to singular (a phenomenon analogous to multicollinearity in ordinary regression) and the optimization algorithm can fail to converge or appear to converge but to a sub-optimal point. Recentering the X s can help considerably with convergence and with the numerical reliability of estimates. Also, if the SD for different slopes is of different orders of magnitude, rescaling the X s to get similar SDs is desirable.

[See notes on using \mathbf{C} to recenter and 'sphere' \mathbf{T}]

[See notes on freeing the RANDOM model from the fixed MODEL: we can centre the FIXED model and the RANDOM model independently of each other]

7.2.5 Random slopes and variance components parametrization

[SEE NOTES]

7.2.6 Testing hypotheses about T

The obvious way of testing a hypothesis about T is to fit the full model and the restricted model and perform a likelihood ratio test to compare the two models. For example, suppose you want to test whether you need a random slope model with one random X effect or whether a simpler random intercept model would be adequate. Fit both models (with the same fixed effects) and record the **deviance** for each model. The **deviance** is a synonym for **-2 Log Likelihood** and is shown in the ‘Fit Statistics’ in the SAS output. Suppose the random intercept model shows:

Fit Statistics	
-2 Log Likelihood	424.6

and the random slope model shows

Fit Statistics	
-2 Log Likelihood	429.9

Next we need to figure out the difference in the number of parameters. Since we used the same fixed model the only difference in parameters comes from the variances of the random model. The smaller model has:

$$T = [\tau_{00}] \tag{122}$$

The larger model has:

$$T = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix} \tag{123}$$

and, since T is symmetric with $\tau_{10} = \tau_{01}$, the larger model has two more parameters than the smaller model. The usual procedure for likelihood ratio tests calls for taking the difference of the two deviances and comparing it to a Chi-Squared distribution with $q = 2$ degrees of freedom = difference in the number of parameters. So, here, we would have:

$$\chi^2 = 429.9 - 424.6 = 5.3 \tag{124}$$

Using a suitable table for the Chi-Square with two degrees of freedom, we find that the p-value is **0.07065**.

Although this is a very common approach to testing this kind of hypothesis, it is somewhat flawed. The reason is that the hypothesis $\tau_{11} = 0$ is that the *edge* of the parameter space

of the full model. (Technically, the null hypothesis is not in the interior of the full model). The usual asymptotic theory for Likelihood Ratio can not be relied upon.

Intuitively, what happens is that, when $\tau_{11} = 0$, the fit is as likely to show underdispersion as overdispersion for the between cluster slope variance. When there is underdispersion, $\hat{\tau}_{11}$ will be 0. Now, $\hat{\tau}_{11}$ contribute to the value of Chi-Square through $\hat{\tau}_{11}^2$ only when $\hat{\tau}_{11}$ is positive (about 1/2 of the time). The hypothesis $\tau_{10} = 0$ contributes to the Chi-Square through τ_{10}^2 whether $\hat{\tau}_{10}$ is positive or not. As a result the distribution of the test statistic under the null hypothesis can be thought of as being a Chi-Square with 1 degree of freedom half the time and a Chi-Square with 2 degrees of freedom the other half. The standard approach above assumes the Chi-Square has 2 degrees of freedom all of the time.

A Chi-Square with 1 degree of freedom has a 'smaller' distribution than one with 2, so this has the effect of making the *actual* distribution of the statistic smaller than its *reference* distribution. The result is that the p-values computed with the reference distribution will be too large: if the statistic is too small then the probability in the upper tail will be too large. The p-value too large implies that the results are interpreted as less significant than they ought to be, i.e. **the true probability of Type I error is lower than you think it is**. You might think this is fine because that means the test is overly **conservative**. But remember that this also means that the **the true probability of Type II error is higher than it would be if you used an 'exact' test**. A test that is conservative with respect to rejection is liberal with respect to acceptance. This might be seen as a problem since researchers are frequently interested in dropping random slopes to reduce the complexity of the random model.

We can find the correct p-value for the problem of testing whether to add one random slope and all its covariances to a model that already has k random slopes and intercept together with all covariances, i.e. the full model has:

$$T = \begin{bmatrix} & & & \tau_{0,k+1} \\ & T_{kk} & & \vdots \\ & & & \tau_{k,k+1} \\ \tau_{k+1,0} & \cdots & \tau_{k+1,k} & \tau_{k+1,k+1} \end{bmatrix} \quad (125)$$

The restricted model has $T = T_{kk}$ and provides a test of $H_0 \tau_{0,k+1} = \dots = \tau_{k,k+1} = \tau_{k+1,k+1} = 0$. The first k parameters are covariance parameters which can take values greater than or less than 0 but the last parameter, $\tau_{k+1,k+1}$, can only take non-negative values. The Chi-Squared for the likelihood ratio test has a nominal Chi-Squared distribution with $q = k + 1$ degrees of freedom but its real distribution under the null hypothesis is that of a mixture of a Chi-Square with with $k + 1$ degrees of freedom and a Chi-Square with k degrees of freedom. We can adjust for this in two ways. If you have output providing the

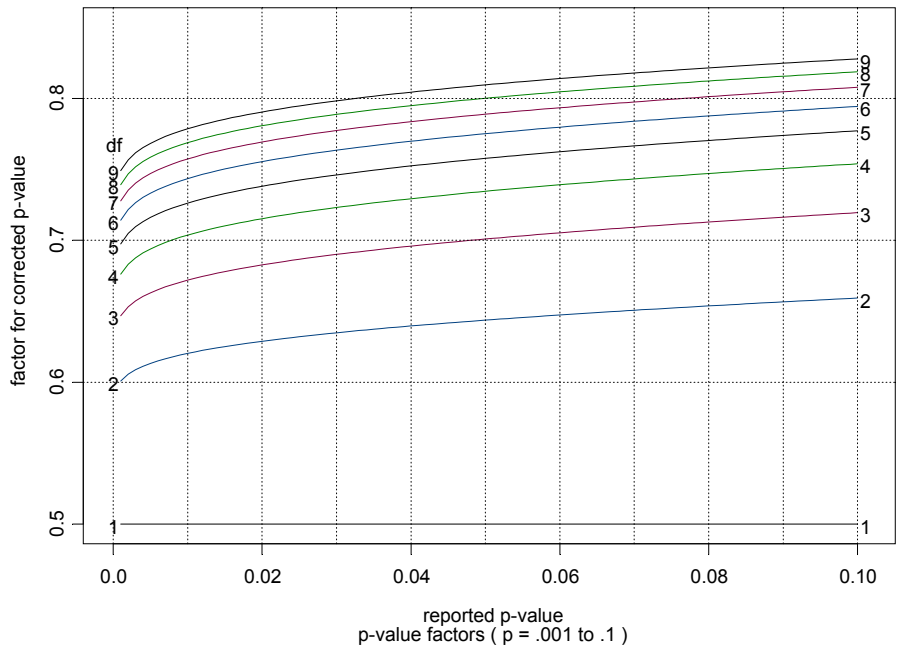


Figure 1: Factor to adjust nominal p-value to get true p-value for various degrees of freedom (p-values from 0.001 to 0.1)

nominal p-value, you can adjust the p-value downward but using the factor shown in Figures 1 and ??.

Alternatively, if you have computed a deviance by taking the difference of two deviances, you can use the table below to find a range for the true p-value:

Critical values for mixture of two Chi-Squares with $df = q$ and $q-1$

df	0.1	0.05	0.01	0.005	0.001	0.0005	0.0001	1e-005
1	1.64	2.71	5.41	6.63	9.55	10.83	13.83	18.19
2	3.81	5.14	8.27	9.63	12.81	14.18	17.37	21.94
3	5.53	7.05	10.50	11.97	15.36	16.80	20.15	24.91
4	7.09	8.76	12.48	14.04	17.61	19.13	22.61	27.54
5	8.57	10.37	14.32	15.97	19.69	21.27	24.88	29.96
6	10.00	11.91	16.07	17.79	21.66	23.29	27.02	32.24

7	11.38	13.40	17.76	19.54	23.55	25.23	29.06	34.41
8	12.74	14.85	19.38	21.23	25.37	27.10	31.03	36.51
9	14.07	16.27	20.97	22.88	27.13	28.91	32.94	38.53

For comparison, this is a table of critical values for the Chi-Square distribution

	0.1	0.05	0.01	0.005	0.001	0.0005	0.0001	1e-005
1	2.71	3.84	6.63	7.88	10.83	12.12	15.14	19.51
2	4.61	5.99	9.21	10.60	13.82	15.20	18.42	23.03
3	6.25	7.81	11.34	12.84	16.27	17.73	21.11	25.90
4	7.78	9.49	13.28	14.86	18.47	20.00	23.51	28.47
5	9.24	11.07	15.09	16.75	20.52	22.11	25.74	30.86
6	10.64	12.59	16.81	18.55	22.46	24.10	27.86	33.11
7	12.02	14.07	18.48	20.28	24.32	26.02	29.88	35.26
8	13.36	15.51	20.09	21.95	26.12	27.87	31.83	37.33
9	14.68	16.92	21.67	23.59	27.88	29.67	33.72	39.34

7.3 Examples

We will illustrate many of the concepts introduced so far with our subsample of the Bryk and Raudenbush data.

First we consider a fixed effects model for the relationship between MATHACH and SES:

SAS input:

```
proc glm data = spida.hs;
  class school;
  model mathach = ses school / solution;
run;
```

Selected SAS output:

The GLM Procedure					
Dependent Variable: MATHACH					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	40	19726.86814	493.17170	13.21	<.0001

Error	1679	62701.74930	37.34470
Corrected Total	1719	82428.61743	

R-Square	Coeff Var	Root MSE	MATHACH Mean
0.239321	46.60822	6.111031	13.11149

Source	DF	Type I SS	Mean Square	F Value	Pr > F
SES	1	11770.23946	11770.23946	315.18	<.0001
SCHOOL	39	7956.62868	204.01612	5.46	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
SES	1	4171.108312	4171.108312	111.69	<.0001
SCHOOL	39	7956.628677	204.016120	5.46	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	13.41999619 B	0.80723095	16.62	<.0001
SES	2.32422573	0.21992118	10.57	<.0001
SCHOOL 1317	-1.04494131 B	1.18939271	-0.88	0.3798
SCHOOL 1374	-3.66214705 B	1.40930069	-2.60	0.0094

.....

We note that the effect of SES is estimated as 2.32. This is the ‘within-school’ effect of SES. It is a weighted average of the individual within-school effects of SES.

7.4 Fitting a multilevel model: contextual effects

In this section we address two problems:

1. To what extent is the effect of SES an individual effect and to what extent is it a school effect? Our models so far have conflated these two effects.
2. Can we estimate the individual effect of SES without contaminating our estimate with the school effect?

We solve both problems at once by introducing contextual effects. We decompose SES into two components:

1. average SES in each school (SES_MEAN) and
2. the deviation of a student from the average SES in the school (SES_ADJ).

There are other ways of creating these two variables and we will discuss them later. The impact of these two variables lies in each variable providing a control for the other. In a model with these two variables, the coefficient of SES_ADJ measure the individual effect keeping the school effect constant. It can be shown that this is the ‘within’ effect of a ‘fixed model’ as the term is used in econometrics. The effect of SES_MEAN keeps SES_ADJ equal to 0, i.e. the school effect compares two schools and two students with constant SES_ADJ which is equivalent to having the individual SES of each student equal to the school SES_MEAN. Thus the effect of SES_MEAN is unconditional while the effect of SES_ADJ is conditional.

We will use SAS to go through the steps of creating SES_MEAN and SES_ADJ.

If nothing is gained by decomposing SES into SES_MEAN and SES_ADJ, the fitted coefficients for SES_MEAN and SES_ADJ will not differ significantly since, in this case, the original SES fits as well as the two separate variables::

$$E(Y_{ij}) = \dots + \beta(X_{ij} - \bar{X}_{.j}) + \beta\bar{X}_{.j} + \dots \quad (126)$$

$$= \dots + \beta X_{ij} + \dots \quad (127)$$

Thus we can use an ESTIMATE statement to test equality. This test is analogous to the test for a random effects model in econometrics.

7.4.1 Example

We will work through an analysis the HS data including a few extra touches to show some capabilities of SAS.

We first define formats for some of the coded variables and summarize the data: These formats allow SAS to show the names of categories in some printed output:

```
proc format; /* formatting variables */
  value sexfmt 1=female 0=male;
  value schfmt 1=Catholic 0=Public;
  value minfmt 1=minority 0=non_minority;
run;
```

The formats are identified with their respective variables:

```

data hs;
  set mixed.hs;
  format female sexfmt.
         minority minfmt.
         sector schfmt. ;
  label ses='social economic status'
         mathach='mathematics achievement';
run;

```

The dataset is sorted by the numerical variable SCHOOL so we can avoid declaring it as a CLASS variable in PROC MIXED:

```

proc sort data=hs;   /* sort by school so we can use it as a numeric var */
  by school;
run;

```

We have a look at the data:

```

proc contents data = hs;
run;

```

```

proc summary data=hs print min q1 median q3 max;
  var school minority female ses mathach size sector meanses;
run;

```

```

proc means data=hs ;
  var ses mathach size  meanses;
  class sector ;
run;

```

```

proc freq data=hs;
  tables sector female minority;
run;

```

We now try a first run with PROC MIXED deliberately using TYPE = UN in the RANDOM statement.

```
ods output SolutionF=fixest;/*Data set'fixest' includes fixed estimates from fitted model
ods output SolutionR=ranest;/*Data set 'ranest' includes random estimates of coefficients
proc mixed data=hs;
  model mathach = ses sector ses*sector /
    outp=predictsch outpm=predictpop solution
    solution ddfm = satterth
    corrb covb covbi cl;

  random int ses /
    sub = school
    type = un
    solution
    G GC GCORR GI
  ;
run;
```

The LOG output gives us a quiet warning:

```
NOTE: Convergence criteria met.
NOTE: Estimated G matrix is not positive definite.
```

The output listing for G shows that SAS is very confused:

Estimated G Matrix				
Row	Effect	Subject	Col1	Col2
1	Intercept	1	3.5693	0.5808
2	SES	1	0.5808	

Estimated Inv(G) Matrix				
Row	Effect	Subject	Col1	Col2
1	Intercept	1		1.7218

2	SES	1	1.7218	-10.5818
---	-----	---	--------	----------

Estimated Chol(G) Matrix				
Row	Effect	Subject	Col1	Col2
1	Intercept	1	1.8893	
2	SES	1	0.3074	

Estimated G Correlation Matrix				
Row	Effect	Subject	Col1	Col2
1	Intercept	1	1.0000	1.0000
2	SES	1	1.0000	1.0000

The blank for the lower right element of the G matrix is supposed to indicate something close to 0 but, if that is the case the matrix is not even a variance matrix. This become clear when we look at the Inv(G) matrix which should also be a variance matrix but has a negative diagonal element. Under these circumstances the Chol(G) matrix and the Correlation matrix should not exist but SAS valiantly reports some numbers for them.

We try again, this time with a better parametrization of the G matrix. We also reverse SES and the INT term in the random statement:

```
proc mixed data=hs;
  model mathach = ses sector ses*sector /
    outp=predictsch outpm=predictpop solution
    solution ddfm=satterth
    corrb covb covbi cl;

  random ses int/
    sub = school
    type = fa0(2)
    solution
    G GC GCORR GI;
run;
```

We now get a legitimate variance matrix for G but it is ‘singular’ i.e. the variation falls entirely on a line. The Choleski root tells us that the point of minimum variance is far from the range of the data.

Estimated G Matrix				
Row	Effect	Subject	Col1	Col2
1	SES	1	0.07119	0.5014
2	Intercept	1	0.5014	3.5309

Estimated Inv(G) Matrix				
Row	Effect	Subject	Col1	Col2
1	SES	1	14.0467	
2	Intercept	1		

Estimated Chol(G) Matrix				
Row	Effect	Subject	Col1	Col2
1	SES	1	0.2668	
2	Intercept	1	1.8791	

Estimated G Correlation Matrix				
Row	Effect	Subject	Col1	Col2
1	SES	1	1.0000	1.0000
2	Intercept	1	1.0000	1.0000

This suggest that the variability in slopes is very small after accounting for heterogeneity accounted for by SECTOR. We refit with only a random intercept.

```
proc mixed data=hs;
  model mathach = ses sector ses*sector /
    outp=predictsch outpm=predictpop
    solution ddfm=satterth
    corrb covb covbi cl;
  random int /sub=school type=fa0(1) solution;
run;
```

We compare the deviances (-2 Res Log Like) of the two models and find an increase of 1.2 for 2 degrees of freedom so we have not lost anything significant by dropping the two variance parameters. Note that we are comparing two models that do not differ in their MODEL statements which allows us to compare deviances even though the models have been fit with the default METHOD = REML.

The table of estimates of fixed effects shows an interesting story ... so far. Everything seems quite significant.

Solution for Fixed Effects						
Effect	Estimate	Standard Error	DF	t Value	Pr > t	Alpha
Intercept	11.6996	0.4285	1225	27.31	<.0001	0.05
SES	2.8917	0.2659	1716	10.88	<.0001	0.05
SECTOR	2.4716	0.7037	1250	3.51	0.0005	0.05
SES*SECTOR	-1.1003	0.4492	1716	-2.45	0.0144	0.05

We will now consider **contextual effects**: why do high SES schools do better? Is it the school or is the kids? So far we haven't made a distinction. Is there an effect of school SES distinct from child SES? Although it doesn't seem likely they could, in principle, even go in opposite directions!

We turn SES into 2 variables: aggregate 'between' school ses and a within school residual or 'school-adjusted ses'

We first create a data set with one observation per school containing the average SES_MEAN for each school.

```
proc means data=hs;
  var ses;
  by school;
  output out=new mean=ses_mean;
```

```

run;

proc print data = new;
run;

```

Just to be on the safe side we sort before merging:

```

proc sort data=new;
  by sschool;
run;

```

We now merge the school averages back into the main data set and we create a variable with each students' deviation, SES_ADJ, in SES from the school average.

```

data hs;
  merge hs new;
  by school;
  ses_adj = ses - ses_mean;
run;

```

We are ready to run PROC MIXED on the new variables. We add the new inner variable SES_ADJ to the RANDOM model.

```

proc mixed data = hs cl method = ml;
  model mathach = ses_adj ses_mean sector ses_adj*sector ses_mean*sector /
    cl ddfm=satterth;
  random ses_adj int / sub = school
    type = fa0(2) G GC GCORR;
run;

```

The G matrix is still close to singular and the point of minimum variance is far outside the range of SES:

Estimated Chol(G) Matrix				
Row	Effect	Subject	Col1	Col2
1	ses_adj	1	0.1351	
2	Intercept	1	1.3193	0.02311

Estimated G Correlation Matrix				
Row	Effect	Subject	Col1	Col2
1	ses_adj	1	1.0000	0.9998
2	Intercept	1	0.9998	1.0000

We refit with a random intercept and get the following estimates of fixed effects:

Effect	Estimate	Standard Error	DF	t Value	Pr > t	Alpha
Intercept	11.7392	0.3511	1130	33.44	<.0001	0.05
ses_adj	2.6815	0.2736	1672	9.80	<.0001	0.05
ses_mean	6.5073	0.9292	1115	7.00	<.0001	0.05
SECTOR	1.4300	0.7730	1044	1.85	0.0646	0.05
ses_adj*SECTOR	-1.0102	0.4600	1672	-2.20	0.0282	0.05
ses_mean*SECTOR	-1.9568	1.7190	1022	-1.14	0.2552	0.05

This output suggests that SES_MEAN and SES_ADJ have very different effects, at least within the Public school system which is the reference level for the coding of SECTOR. There is also a suggestion that SECTOR might not be important in this new model.

We rerun the model with the following CONTRAST statement:

```
contrast 'sector' SECTOR 1, ses_adj*SECTOR 1, ses_mean*SECTOR 1;
```

which allows us to test the hypothesis that all terms involving SECTOR are simultaneously equal to 0. The result is not entirely clear:

	Contrast			
	Num	Den		
Label	DF	DF	F Value	Pr > F
sector	3	1201	2.75	0.0415

Perhaps the apparent interaction between sector and SES might show up as an interaction between SES_MEAN and SES_ADJ suggesting that high SES schools have flatter MATHACH/SES slopes than low SES schools. The apparently greater fairness of Catholic schools in the early analyses might have been due to the fact that they tend to have higher values of SES_MEAN.

Another possible explanation is that failing to separate the within school effect of SES from the between school effect *contaminated* the estimate of the between-school effect with the lower within-school effect. This is, ironically, the reverse of the concern usually expressed in econometrics. Having a lower coefficient than it should, SES, as a proxy for SES_MEAN, in the between-school model underadjusted for the effect of SES and left variation to be explained by SECTOR. Once SES_MEAN was uncoupled with SES_ADJ, it accounted for the variation that had been attributed to SECTOR.

The following questions are left as exercises:

1. Use the ESTIMATE or CONTRAST statement to carry out a formal test of the equality of the coefficients of SES_MEAN and SES_ADJ. This test is an analogue of the Wu-Hausman test in econometrics.
2. Explore the usefulness of a SES_MEAN*SES_ADJ interaction. How would you interpret it?

7.5 Diagnostics

Diagnostic for multilevel models is an area in which there is much room for developments. Few books put much stress on diagnostics with the exception of Pinheiro and Bates (2000).

There are some obvious applications of the diagnostics developed for regression Fox (1991). For articles specifically on diagnostics for mixed and longitudinal models consult Hodges (1998), Banerjee and Frees (1997), Bradlow and Zaslavsky (1997), Albert and Chib (1997), Christensen and Bedrick (1997), Ho et al. (1995), Christensen et al. (1992), Ahn (1990), Hocking et al. (1989), Wernecke et al. (1988), Beckman et al. (1987), Hocking and Bremer (1987).

8 Longitudinal Data

The structure of longitudinal data shares much with that of multilevel data. Turn the school into a subject and the students in the school into times or ‘occasions’ on which the subject is measured and, except for the fact that measurements are ordered in time, the structures are essentially the same. Mixed models offer only one of many ways of approaching the analysis of longitudinal data but, in many ways, mixed models are extensions of some traditional ways and provide more flexibility. We will briefly review traditional methods and then discuss the use of mixed models.

The best method depends on a number of factors:

- The number of subjects: if the number is very small it might not be feasible to treat subjects as random. The analysis then focuses on an ‘anecdotal’ description of the observed subject and does not generalize to the population of subjects. Statistical models play a role in analyzing the structure of responses within subjects especially if the subjects were measured on a large number of occasions. This situation is typical or research in psychotherapy where a few subjects are observed over a long time. With long series and few subjects, **time series analysis** can be an appropriate method.
- The number of occasions and whether occasions are fixed from subject to subject or variable. With a small number (<10?) of fixed occasions and with the same ‘within subject’ design for each subject, the traditional methods of *repeated measures analysis* can be used. There are two common repeated measures models: **univariate repeated measures** and **multivariate repeated measures**. In univariate repeated measures models the only random variability from subject to subject comes from a random intercept. In multivariate repeated measures, no constraints are placed on the intercorrelations between occasions. These two models can be thought of as extreme cases: a model with only one covariance parameters and a model with all possible covariance parameters. These two models turn out to be special cases of **mixed models** which provide an approach with which we can specify the whole range of models between these extremes, models that allow simpler covariance structures based on the structure of the data and the fact that, having been observed in time, observations that are close in time are often expected to be more strongly correlated than observations that are farther in time.
- **Mixed model** methods are particularly well suited to data with many subjects and variable within-subject designs, either due to variable occasions or to time-varying covariates.

We will look at five data examples to illustrate the breadth of applications of mixed models to longitudinal data:

- The classical Potthoff and Roy (1964) data of jaw sizes of ?? children at ages 8, 10, 12 and 14. This is a simple yet interesting example of longitudinal data and helps understand the major concepts in a simple context.
- PSID: Panel Study of Income Dynamics: An excerpt from the PSID conducted by ISR at the University of Michigan at Ann Arbor will illustrate the application of methods on a survey data set with fixed occasions and time-varying covariates.
- IQ recovery of post-coma patients: Long-term non-linear model with highly variable occasions.
- Migraine diaries: A dichotomous dependent variable with a non-linear long-term response model, as seasonal periodic component and a long-term non-stationarity.

8.1 The basic model

We can think of longitudinal data as multilevel data with t (for a time index) replacing the within cluster index i . We could keep j for subjects but to be consistent with Snijders and Bosker (1999) we will switch from j to i . The other difference is that at least one of the X variables will be time or a function of time. A simple model for the Potthoff and Roy data with a linear trend in time that varies randomly from child to child would look just like a multilevel model with the following level 1 model:

$$Y_{it} = \beta_{0i} + \beta_{1i}X_{it} + \varepsilon_{it} \quad (128)$$

where i ranges over the 27 subjects, t is the time index: $t = 1, 2, 3, 4$, and X_{ti} takes on the actual years : 8, 10, 12, 14. Now we come to the main departure from multilevel models: what to assume about ε_{ti} . It no longer seems reasonable to unquestioningly assume that ε_{ti} are independent as t varies for a given i . Observations close in time might be more strongly related than observations that are far apart. (Note that this relationship does not necessarily lead to a positive correlation.)

The Laird-Ware form of the model for a single child has the form:

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i \quad (129)$$

$$= \mathbf{X}_i\boldsymbol{\gamma} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\epsilon}_i \quad (130)$$

with $\mathbf{X}_i = \mathbf{Z}_i = \mathbf{X}$. With the matrices written out, we get:

$$\begin{bmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ Y_{i4} \end{bmatrix} = \begin{bmatrix} 1 & X_{i1} \\ 1 & X_{i2} \\ 1 & X_{i3} \\ 1 & X_{i4} \end{bmatrix} \begin{bmatrix} \beta_{0i} \\ \beta_{1i} \end{bmatrix} + \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \end{bmatrix} \quad (131)$$

$$= \begin{bmatrix} 1 & X_{i1} \\ 1 & X_{i2} \\ 1 & X_{i3} \\ 1 & X_{i4} \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \end{bmatrix} + \begin{bmatrix} 1 & X_{i1} \\ 1 & X_{i2} \\ 1 & X_{i3} \\ 1 & X_{i4} \end{bmatrix} \begin{bmatrix} u_{0i} \\ u_{1i} \end{bmatrix} + \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \end{bmatrix} \quad (132)$$

with

$$\text{Var} \left(\begin{bmatrix} u_{0i} \\ u_{1i} \end{bmatrix} \right) = \mathbf{T} = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix} \quad (133)$$

With a standard multilevel model we would have:

$$\Sigma = \text{Var} \left(\begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \varepsilon_{3i} \\ \varepsilon_{4i} \end{bmatrix} \right) = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix} \quad (134)$$

but this is often not realistic with data ordered in time. A model that allows autocorrelations to have a form called AR(1) or ‘auto-regressive of order 1’ would have:

$$\Sigma = \text{Var} \left(\begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \varepsilon_{3i} \\ \varepsilon_{4i} \end{bmatrix} \right) = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix} \quad (135)$$

Here we need to be careful because our model is identified only through the marginal variance:

$$\mathbf{V} = \mathbf{Z}\mathbf{T}\mathbf{Z}' + \Sigma \quad (136)$$

We now have 5 variance parameters in the model and 10 variance estimates. It might seem that identifiability should not be a problem but we need to be on guard for possible collinearities. We can’t add arbitrary complexity to Σ without thinking that the estimation of the parameters in both \mathbf{T} and Σ take place together. Writing out the \mathbf{V} matrix we see that we don’t have 10 independent variance estimates. If we centre the time variable, we

get:

$$\begin{aligned}
\mathbf{V} &= \begin{bmatrix} 1 & -3 \\ 1 & -1 \\ 1 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ -3 & -1 & 1 & 3 \end{bmatrix} + \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix} \\
&= \tau_{00} \mathbf{U} + \begin{bmatrix} -6\tau_{01} + 9\tau_{11} + \sigma^2 & -4\tau_{01} + 3\tau_{11} + \sigma^2\rho & -2\tau_{01} - 3\tau_{11} + \sigma^2\rho^2 & -9\tau_{11} + \sigma^2\rho^3 \\ & -2\tau_{01} + \tau_{11} + \sigma^2 & -\tau_{11} + \sigma^2\rho & 2\tau_{01} - 3\tau_{11} + \sigma^2\rho^2 \\ & & 2\tau_{01} + \tau_{11} + \sigma^2 & 4\tau_{01} + 3\tau_{11} + \sigma^2\rho \\ & & & 6\tau_{01} + 9\tau_{11} + \sigma^2 \end{bmatrix} \quad (137)
\end{aligned}$$

so that the upper triangle of the variance matrix is

$$\begin{bmatrix} v_{11} \\ v_{12} \\ v_{13} \\ v_{14} \\ v_{22} \\ v_{23} \\ v_{24} \\ v_{33} \\ v_{34} \\ v_{44} \end{bmatrix} = \begin{bmatrix} \tau_{00} - 6\tau_{01} + 9\tau_{11} + \sigma^2 \\ \tau_{00} - 4\tau_{01} + 3\tau_{11} + \sigma^2\rho \\ \tau_{00} - 2\tau_{01} - 3\tau_{11} + \sigma^2\rho^2 \\ \tau_{00} - 9\tau_{11} + \sigma^2\rho^3 \\ \tau_{00} - 2\tau_{01} + \tau_{11} + \sigma^2 \\ \tau_{00} - \tau_{11} + \sigma^2\rho \\ \tau_{00} + 2\tau_{01} - 3\tau_{11} + \sigma^2\rho^2 \\ \tau_{00} + 2\tau_{01} + \tau_{11} + \sigma^2 \\ \tau_{00} + 4\tau_{01} + 3\tau_{11} + \sigma^2\rho \\ \tau_{00} + 6\tau_{01} + 9\tau_{11} + \sigma^2 \end{bmatrix} \quad (139)$$

Heuristically, to estimate the variance parameters, we need to estimate \mathbf{V} and then backsolve for a set of variance parameters that reproduce \mathbf{V} . The transformation is not linear since it involves different powers of ρ but we can consider the Jacobian:

$$\frac{\partial \mathbf{V}}{\partial \Phi} = \frac{\partial}{\partial \Phi} \begin{bmatrix} v_{11} \\ v_{12} \\ v_{13} \\ v_{14} \\ v_{22} \\ v_{23} \\ v_{24} \\ v_{33} \\ v_{34} \\ v_{44} \end{bmatrix} = \begin{bmatrix} 1 & -6 & 9 & 1 & 0 \\ 1 & -4 & 3 & \rho & 1 \\ 1 & -2 & -3 & \rho^2 & 2\rho \\ 1 & 0 & -9 & \rho^3 & 3\rho^2 \\ 1 & -2 & 1 & 1 & 0 \\ 1 & 0 & -1 & \rho & 1 \\ 1 & 2 & -3 & \rho^2 & 2\rho \\ 1 & 2 & 1 & 1 & 0 \\ 1 & 4 & 3 & \rho & 1 \\ 1 & 6 & 9 & 1 & 0 \end{bmatrix} \quad \text{where } \Phi = [\tau_{00} \quad \tau_{01} \quad \tau_{11} \quad \sigma^2 \quad \rho] \quad (140)$$

The information on Φ is related to the Gramian matrix $\mathbf{J}'\mathbf{J}$ of the Jacobian matrix \mathbf{J} whose normalized form has a condition index of over 1,000 for values of $\rho > .62$.

As mentioned earlier, both univariate and multivariate repeated measures can be represented as ‘extreme’ cases of mixed models. Univariate repeated measures is equivalent to a random intercept model which generates a ‘compound symmetric’ \mathbf{V} matrix: In this model \mathbf{X} generates a model for time, generally a saturated model (i.e. equivalent to a categorical model) for time. \mathbf{Z} consists of a column of 1s. Using a cubic polynomial model for the centered Pothoff and Roy data:

$$\begin{bmatrix} Y_{1i} \\ Y_{2i} \\ Y_{3i} \\ Y_{4i} \end{bmatrix} = \begin{bmatrix} 1 & -3 & 9 & -27 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & 3 & 9 & 27 \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} [u_{0i}] + \begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \varepsilon_{3i} \\ \varepsilon_{4i} \end{bmatrix} \quad (141)$$

with $\text{Var}(u_{0i}) = \tau_{00}$ and $\text{Var}(\varepsilon_i) = \sigma^2\mathbf{I}$. Thus

$$\mathbf{V} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} [\tau_{00}] \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix} + \begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix} \quad (142)$$

$$= \begin{bmatrix} \tau_{00} + \sigma^2 & \tau_{00} & \tau_{00} & \tau_{00} \\ \tau_{00} & \tau_{00} + \sigma^2 & \tau_{00} & \tau_{00} \\ \tau_{00} & \tau_{00} & \tau_{00} + \sigma^2 & \tau_{00} \\ \tau_{00} & \tau_{00} & \tau_{00} & \tau_{00} + \sigma^2 \end{bmatrix} \quad (143)$$

A drawback of this model is that it assumes the same covariance between pairs of observations whether they are close or far in time. Note that the random slopes model (without autocorrelation) may seem better in this regard. Setting $\tau_{01} = 0$ which is the same as

assuming that the point of minimum variance is at the origin:

$$\mathbf{V} = \begin{bmatrix} 1 & -3 \\ 1 & -1 \\ 1 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ -3 & -1 & 1 & 3 \end{bmatrix} + \begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix} \quad (144)$$

$$= \tau_{00}\mathbf{U} + \quad (145)$$

$$\begin{bmatrix} -6\tau_{01} + 9\tau_{11} + \sigma^2 & -4\tau_{01} + 3\tau_{11} & -2\tau_{01} - 3\tau_{11} & -9\tau_{11} \\ & -2\tau_{01} + \tau_{11} + \sigma^2 & -\tau_{11} & 2\tau_{01} - 3\tau_{11} \\ & & 2\tau_{01} + \tau_{11} + \sigma^2 & 4\tau_{01} + 3\tau_{11} \\ & & & 6\tau_{01} + 9\tau_{11} + \sigma^2 \end{bmatrix} \quad (146)$$

$$= \begin{bmatrix} \tau_{00} + 9\tau_{11} + \sigma^2 & \tau_{00} + 3\tau_{11} & \tau_{00} - 3\tau_{11} & \tau_{00} - 9\tau_{11} \\ & \tau_{00} + \tau_{11} + \sigma^2 & \tau_{00} - \tau_{11} & \tau_{00} - 3\tau_{11} \\ & & \tau_{00} + \tau_{11} + \sigma^2 & \tau_{00} + 3\tau_{11} \\ & & & \tau_{00} + 9\tau_{11} + \sigma^2 \end{bmatrix} \quad (147)$$

Suitable values of τ_{00} , τ_{11} and σ^2 will approximate an AR(1) correlation structure. Using a random intercept and and AR(1) parameter uses one fewer parameters than (137) and yields better identification for ρ :

$$\mathbf{V} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} [\tau_{00}] \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix} + \begin{bmatrix} \sigma^2 & \rho & \rho^2 & \rho^3 \\ \rho & \sigma^2 & \rho & \rho^2 \\ \rho^2 & \rho & \sigma^2 & \rho \\ \rho^3 & \rho^2 & \rho & \sigma^2 \end{bmatrix} \quad (148)$$

The MANOVA repeated measures design is saturated for \mathbf{V} . It can be represented as a mixed models through \mathbf{T} or through $\mathbf{\Sigma}$. For example if \mathbf{X} yields a saturated model, we could have:

$$\mathbf{Y}_i = \mathbf{X}\boldsymbol{\gamma} + \mathbf{Z}\mathbf{u}_i + \boldsymbol{\varepsilon}_i \quad (149)$$

$$\begin{bmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ Y_{i4} \end{bmatrix} = \begin{bmatrix} 1 & -3 & 9 & -27 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & 3 & 9 & 27 \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix} + \begin{bmatrix} 1 & -3 & 9 & -27 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & 3 & 9 & 27 \end{bmatrix} \begin{bmatrix} u_{0i} \\ u_{1i} \\ u_{2i} \\ u_{3i} \end{bmatrix} + \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \end{bmatrix} \quad (150)$$

which yields

$$\text{Var} \left(\begin{bmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ Y_{i4} \end{bmatrix} \right) = \begin{bmatrix} 1 & -3 & 9 & -27 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & 3 & 9 & 27 \end{bmatrix} \begin{bmatrix} \tau_{00} & \tau_{01} & \tau_{02} & \tau_{03} \\ \tau_{10} & \tau_{11} & \tau_{12} & \tau_{13} \\ \tau_{20} & \tau_{21} & \tau_{22} & \tau_{23} \\ \tau_{30} & \tau_{31} & \tau_{32} & \tau_{33} \end{bmatrix} \begin{bmatrix} 1 & -3 & 9 & -27 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & 3 & 9 & 27 \end{bmatrix}' \quad (151)$$

$$+ \begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix} \quad (152)$$

$$\mathbf{V} = \mathbf{\Phi} + \sigma^2 \mathbf{I} \quad (153)$$

where $\mathbf{\Phi}$ is free to be any variance matrix. However, σ^2 is not identifiable and fitting this model using software for mixed models requires the ability to constrain σ^2 to be equal to 0 or some very small quantity. [CHECK: this is possible in NLME and in MIWin but is it possible in SAS PROC MIXED?]

The other approach is to use a saturated model for $\mathbf{\Sigma}$ and no random part. [Possible in MIWin but not in NLME. How about SAS?]

$$\mathbf{Y}_i = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_i \quad (154)$$

$$\begin{bmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ Y_{i4} \end{bmatrix} = \begin{bmatrix} 1 & -3 & 9 & -27 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & 3 & 9 & 27 \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \end{bmatrix} \quad (155)$$

yielding

$$\mathbf{V} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{bmatrix} = \mathbf{\Sigma} \quad (156)$$

which is equivalent in form and substance to MANOVA repeated measures.

We can therefore think of mixed models as extensions of univariate and multivariate repeated measures allowing models with variance matrices of intermediate complexity. They also allow one to fit univariate and multivariate repeated measures models to incomplete data provided the missing cases are ‘missing at random.’

We fit these models by specifying both the $\mathbf{\Phi}$ and the $\mathbf{\Sigma}$ matrices. In SAS the structure of $\mathbf{\Sigma}$ matrix is controlled by the `REPEATED` statement. The options are listed in SAS Institute

Inc. (2000a) on page 1996 and a summary can be found below on page ???. Some of the most important options to specify the *TYPE* of covariance structure are:

Structure	Description	Parms	(i, j) th element
AR(1)	Autoregressive(1)	2	$\sigma^2 \rho^{ i-j }$
ARMA(1,1)	ARMA(1)	3	$\sigma^2 [\gamma \rho^{ i-j } (1 - \delta_{ij}) + \delta_{ij}]$
ARH(1)	Heterog. AR(1)	$T + 1$	$\sigma_i \sigma_j \rho^{ i-j }$
ANTE(1)	Ante-dependence	$2T - 1$	$\sigma_i \sigma_j \prod_{k=i}^{j-1} \rho_k$
FA0(T)	Choleski unstructured	$T(T + 1)/2$	$\sum_{k=1}^{\min(i,j)} \lambda_{ik} \lambda_{jk}$
UN	Unstructured	$T(T + 1)/2$	σ_{ij}
CS	Compound Symmetry	2	$\sigma_1^2 + \tau \delta_{ij}$

Note that δ_{ij} represents the ‘Kronecker delta’ which is equal to 1 when $i = j$ and 0 if $i \neq j$.

AR(1) is used when the dependence over time follows an ‘auto-regressive process of order 1’ which means that the random error for each observation is equal to a constant times the previous error plus a new error (innovation). ARMA combines an autoregressive process with a ‘moving average’ process. It can be used to model situations where the observed error is composed of an underlying AR(1) process plus independent errors. ARH(1) is similar to AR(1) except that it allows observations on each occasion to have different variances. The correlation structure is identical to AR(1). ANTE(1) is similar to ARH(1) except that the ‘autocorrelation’ between adjoining occasions can vary from one adjoining pair to another. The autocorrelation between pairs that are more than one unit of time apart is determined by the autocorrelations between intermediate adjoining pairs.

FA0(T) where T is the number of occasions per subject produces a Choleski parametrization which is superior to an unstructured (UN) since it imposes a positive-semi definite structure for Σ . Unfortunately this approach only works when the number of occasions is the same for each subject.

CS produces a Σ matrix with a compound symmetry structure which is (almost) identical to the matrix obtained with a random intercept model. One subtle difference is that CS can allow negative covariances while these would not be possible under SAS’s positive constraints for variance with a random intercept model.

The covariance structures for some of the types listed above:

AR(1)	σ^2	1	ρ	ρ^2	ρ^3
		ρ	1	ρ	ρ^2
		ρ^2	ρ	1	ρ
		ρ^3	ρ^2	ρ	1
ARMA(1,1)	σ^2	1	γ	$\gamma\rho$	$\gamma\rho^2$
		γ	1	γ	$\gamma\rho$
		$\gamma\rho$	γ	1	γ
		$\gamma\rho^2$	$\gamma\rho$	γ	1
ARH(1)	σ_1^2	$\sigma_1\sigma_2\rho$	$\sigma_1\sigma_3\rho^2$	$\sigma_1\sigma_4\rho^3$	
	$\sigma_2\sigma_1\rho$	σ_2^2	$\sigma_2\sigma_3\rho$	$\sigma_2\sigma_4\rho^2$	
	$\sigma_3\sigma_1\rho^2$	$\sigma_3\sigma_2\rho$	σ_3^2	$\sigma_3\sigma_4\rho$	
	$\sigma_4\sigma_1\rho^3$	$\sigma_4\sigma_2\rho^2$	$\sigma_4\sigma_3\rho$	σ_4^2	

8.2 Analyzing longitudinal data

8.2.1 Classical or Mixed models

A comment from a SAS FAQ comparing PROC MIXED with PROC GLM:

You can use PROC GLM or PROC MIXED in SAS to perform repeated measures ANOVA. Each procedure has strengths and weaknesses; one nice MIXED feature is its ability to perform comparisons involving within and between-subjects factors in the same contrast. PROC GLM uses separate matrices for between-subjects effects versus within-subjects effects. This is not a problem if you are interested in between-subjects effects or within-subjects effects, but it can present complications if you want to generate contrasts that cross levels of both between and within-subjects effects simultaneously. For this reason, this FAQ illustrates the more general contrast approach offered by PROC MIXED.³

One problem in using both approaches is that the input data must be in a different form for each. An FAQ addresses the transformation from the ‘classical’ form to that needed for PROC MIXED:

Question:

I’ve run a repeated measures ANOVA using the SAS GLM procedure. My dataset has a single between-subjects grouping factor with two levels and I have four

³From a SAS FAQ at the University of Texas: <http://www.utexas.edu/cc/faqs/stat/sas/sas94.html>

dependent variables that comprise my repeated measures effect. For follow-up analyses using the MIXED procedure, I need to rearrange my dataset so that there is a single dependent variable column and then a second column that refers to the measurement occasion of the dependent variable (e.g., 1, 2, 3, or 4). What's the best way for me to rearrange my dataset using SAS?

Answer:

There are a number of ways you can SAS to transform your data from multivariate to univariate form. Here is one approach you can use.

```
DATA one ;
INFILE cards ;
INPUT a b1 b2 b3 b4 ;
subjid + 1 ;
CARDS ;
1 3 4 7 7
1 6 5 8 8
1 3 4 7 9
1 3 3 6 8
2 1 2 5 10
2 2 3 6 10
2 2 4 5 9
2 2 3 6 11
;
RUN ;
PROC SORT DATA = one ;
BY a subjid ;
RUN ;
PROC TRANSPOSE DATA=one OUT=two NAME=measure PREFIX=y_all_ ;
VAR b1-b4 ;
BY a subjid ;
RUN ;
```

This example features a between-subjects grouping factor "a" and four within-subjects dependent variables, b1 through b4. The SAS syntax shown above first creates a counter variable called "subjid" to denote subject ID number using the syntax SUBJID+1. The program then sorts the dataset by subject ID number within each level of group using the SORT procedure. The TRANSPOSE

procedure then produces a single dependent variable, "y_all_1", a character variable called "measure" that indicates the measurement occasion of the dependent variable, and it retains the correct group specification "a". Notice that the TRANSPOSE procedure appends a number to the dependent variable name, which is specified by the PREFIX keyword. The prefix "y_all_" is joined to the number "1" to create the new dependent variable, "y_all_1". The new single dependent variable, the measure variable, and the group variable "a" are written to the temporary SAS dataset work.two by the TRANSPOSE procedure.

8.3 Pothoff and Roy

A good illustration of a simple longitudinal model is provided by data in Pothoff and Roy (1964) consisting of measurements on the growth of the jaw for 11 girls and 16 boys at ages 8, 10, 12, and 14. Since the data are complete and balanced, we can choose from many methods for their analysis. The following shows a SAS data step creating the data set. The data are entered in *wide* form (one row per subject) and then reshaped into *long* form (one row per occasion). In general this process is more complicated because all time-varying variables have to be treated like the dependent y variables in the example. (STATA has a function that does this very easily).

```
data prwide;
  input Person Gender $ y1 y2 y3 y4;
  datalines;
  1   F   21.0   20.0   21.5   23.0
  2   F   21.0   21.5   24.0   25.5
  3   F   20.5   24.0   24.5   26.0
  4   F   23.5   24.5   25.0   26.5
  5   F   21.5   23.0   22.5   23.5
  6   F   20.0   21.0   21.0   22.5
  7   F   21.5   22.5   23.0   25.0
  8   F   23.0   23.0   23.5   24.0
  9   F   20.0   21.0   22.0   21.5
 10  F   16.5   19.0   19.0   19.5
 11  F   24.5   25.0   28.0   28.0
 12  M   26.0   25.0   29.0   31.0
 13  M   21.5   22.5   23.0   26.5
 14  M   23.0   22.5   24.0   27.5
```

```

15  M  25.5  27.5  26.5  27.0
16  M  20.0  23.5  22.5  26.0
17  M  24.5  25.5  27.0  28.5
18  M  22.0  22.0  24.5  26.5
19  M  24.0  21.5  24.5  25.5
20  M  23.0  20.5  31.0  26.0
21  M  27.5  28.0  31.0  31.5
22  M  23.0  23.0  23.5  25.0
23  M  21.5  23.5  24.0  28.0
24  M  17.0  24.5  26.0  29.5
25  M  22.5  25.5  25.5  26.0
26  M  23.0  24.5  26.0  30.0
27  M  22.0  21.5  23.5  25.0
;

```

```

data pr; /* converting wide to long */
  set prwide;
  y=y1; Age=8; output;
  y=y2; Age=10; output;
  y=y3; Age=12; output;
  y=y4; Age=14; output;
  drop y1-y4;
run;

```

Perhaps the best way to visualize the data is to plot Y against Age for each subject (Figure 2). The plot immediately reveals at least one anomaly but we will not exclude it in the analysis.

We will consider five models to this data which we will use to estimate two features of the growth process:

1. the difference in the rate of growth between boys and girls and
2. the difference in size at age 14.

The four models we will use are:

1. Univariate ANOVA repeated measures

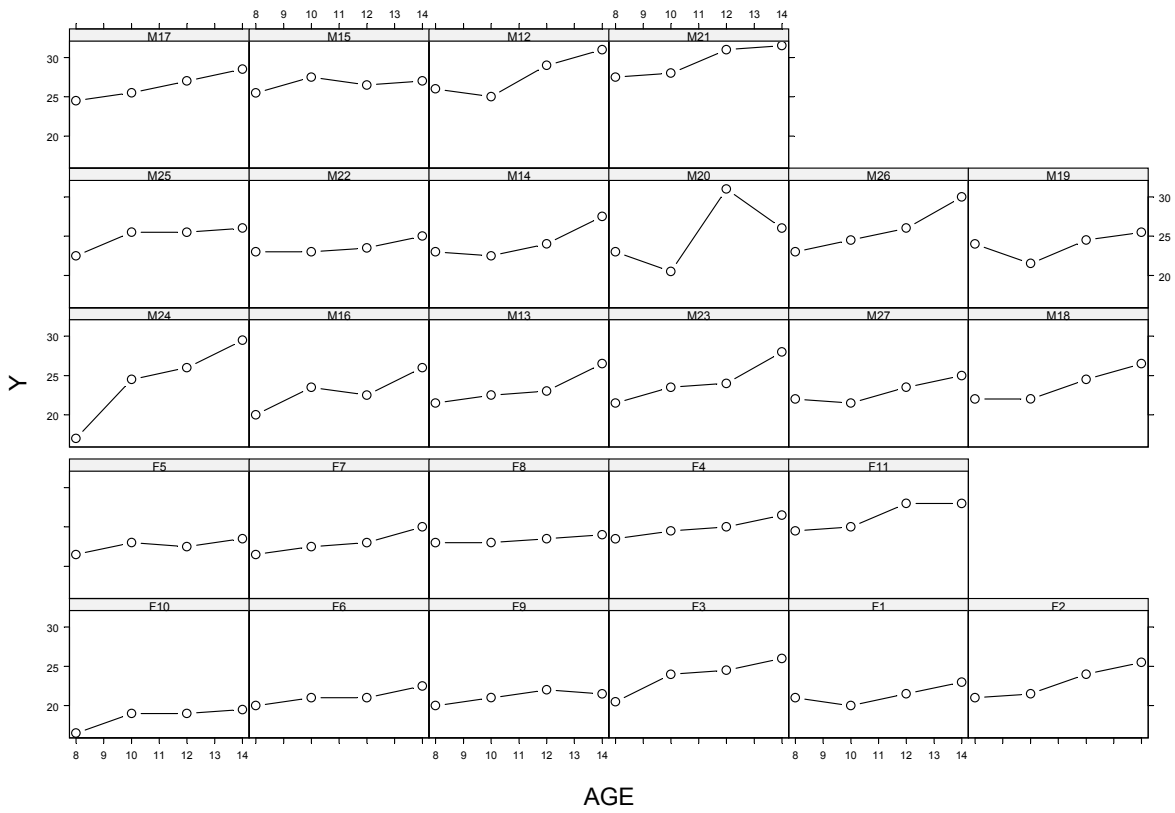


Figure 2: Pothoff and Roy data on the growth of the pterigomandibular fissure of 11 girls and 16 boys between the ages 8 to 14.

2. MANOVA repeated measures
3. Random intercept with auto-correlation
4. Random slope model without auto-correlation
5. Random slope model with auto-correlation

Actually, the last two are exercises!

8.3.1 Univariate ANOVA

The univariate ANOVA repeated measures analysis is a random intercept analysis. We can fit an arbitrary growth curve by fitting a polynomial of degree one less than the number of time periods. If the higher order coefficients are not significant we drop them and fit a linear model. We will assume that this work has already been done and fitting a linear model is adequate: The random intercept model is specified as follows.

Note the construction of the ESTIMATE statement when *Gender* is used as a CLASS variable:

Variable	INT	Gender=F	Gender=M	Age	Gender=F*Age	Gender=M*Age
Male at Age 14	1	0	1	14	0	14
Female at Age 14	1	1	0	14	14	0
Difference (F-M)	0	1	-1	0	14	-14

PROC MIXED generates 2 dummy variables, one for Females and one for Males. These variables are collinear with the INTERCEPT, so when it comes to estimation (see *Solution for Fixed Effects* below) the estimated coefficient for the second dummy variable is forced to 0. However, we must take both dummy variables into account when we build ESTIMATE and CONTRAST statements.

```
proc mixed data=pr method=ml covtest;
  class Person Gender;
  model y = Gender Age Gender*Age / s;
  random intercept / type=fa0(1) subject=Person
    g gc v vcorr;
  estimate 'gap ap 14' Gender 1 -1 Age*Gender 14 -14;
run;
```

Partial output:

Estimated G Matrix			
Row	Effect	Person	Col1
1	Intercept	1	3.0306

Estimated Chol(G) Matrix			
Row	Effect	Person	Col1
1	Intercept	1	1.7409

Covariance Parameter Estimates					
Cov Parm	Subject	Estimate	Standard Error	Z	Pr > Z
FA(1,1)	Person	1.7409	0.2744	6.35	<.0001
Residual		1.8746	0.2946	6.36	<.0001

Fit Statistics	
-2 Log Likelihood	428.6
AIC (smaller is better)	440.6
AICC (smaller is better)	441.5
BIC (smaller is better)	448.4

Estimated V Matrix for Person 1				
Row	Col1	Col2	Col3	Col4
1	4.9052	3.0306	3.0306	3.0306
2	3.0306	4.9052	3.0306	3.0306
3	3.0306	3.0306	4.9052	3.0306
4	3.0306	3.0306	3.0306	4.9052

Estimated V Correlation Matrix for Person 1				
Row	Col1	Col2	Col3	Col4
1	1.0000	0.6178	0.6178	0.6178
2	0.6178	1.0000	0.6178	0.6178
3	0.6178	0.6178	1.0000	0.6178
4	0.6178	0.6178	0.6178	1.0000

Null Model Likelihood Ratio Test		
DF	Chi-Square	Pr > ChiSq
1	49.60	<.0001

Solution for Fixed Effects						
Effect	Gender	Estimate	Standard Error	DF	t Value	Pr > t
Intercept		16.3406	0.9631	25	16.97	<.0001
Gender	F	1.0321	1.5089	79	0.68	0.4960
Gender	M	0
Age		0.7844	0.07654	79	10.25	<.0001
Age*Gender	F	-0.3048	0.1199	79	-2.54	0.0130
Age*Gender	M	0

Label	Estimate	Estimate Standard Error	DF	t Value	Pr > t
gap ap 14	-3.2355	0.8162	79	-3.96	0.0002

To test the hypothesis that the two genders have the same slope we consider the p-value for the interaction between Age and Gender (0.0130) and to test the difference in sizes at age 14, we use the ESTIMATE output showing a difference of -3.2355 with a p-value of 0.0002. Note the variance matrix for person 1. It has constants off diagonal. The correlation matrix has a constant value off-diagonal.

8.3.2 MANOVA repeated measures

The MANOVA repeated measures analysis allows and arbitrary variances and covariances. We could build such a matrix through \mathbf{T} but we would have to force σ^2 to be zero. Instead we can use the generalization of the mixed model:

$$\begin{aligned} \mathbf{Y}_i &= \mathbf{X}\boldsymbol{\gamma} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i \\ \mathbf{u}_i &\sim N(\mathbf{0}, \mathbf{T}) \\ \boldsymbol{\varepsilon}_i &\sim N(\mathbf{0}, \boldsymbol{\Sigma}) \end{aligned}$$

and drop the RANDOM part (\mathbf{Z}_i) but use the REPEATED statement to parametrize $\boldsymbol{\Sigma}$ so that it can be any variance matrix. The code is;

```
proc mixed data=pr method=ml covtest;
  class Person Gender;
```

```

model y = Gender Age Gender*Age / s;
repeated / type=un subject=Person r rcorr;
estimate 'gap at 14' Gender 1 -1 Age*Gender 14 -14;
run;

```

Estimated R Matrix for Person 1

Row	Col1	Col2	Col3	Col4
1	5.1192	2.4409	3.6105	2.5222
2	2.4409	3.9279	2.7175	3.0624
3	3.6105	2.7175	5.9798	3.8235
4	2.5222	3.0624	3.8235	4.6180

Estimated R Correlation Matrix for Person 1

Row	Col1	Col2	Col3	Col4
1	1.0000	0.5443	0.6526	0.5188
2	0.5443	1.0000	0.5607	0.7190
3	0.6526	0.5607	1.0000	0.7276
4	0.5188	0.7190	0.7276	1.0000

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z	Pr > Z
UN(1,1)	Person	5.1192	1.4169	3.61	0.0002
UN(2,1)	Person	2.4409	0.9835	2.48	0.0131
UN(2,2)	Person	3.9279	1.0824	3.63	0.0001
UN(3,1)	Person	3.6105	1.2767	2.83	0.0047
UN(3,2)	Person	2.7175	1.0740	2.53	0.0114
UN(3,3)	Person	5.9798	1.6279	3.67	0.0001
UN(4,1)	Person	2.5222	1.0649	2.37	0.0179
UN(4,2)	Person	3.0624	1.0135	3.02	0.0025
UN(4,3)	Person	3.8235	1.2508	3.06	0.0022
UN(4,4)	Person	4.6180	1.2573	3.67	0.0001

Fit Statistics

-2 Log Likelihood	419.5
-------------------	-------

AIC (smaller is better)	447.5
AICC (smaller is better)	452.0
BIC (smaller is better)	465.6

Null Model Likelihood Ratio Test		
DF	Chi-Square	Pr > ChiSq
9	58.76	<.0001

Solution for Fixed Effects

			Standard			
Effect	Gender	Estimate	Error	DF	t Value	Pr > t

Intercept		15.8423	0.9356	25	16.93	<.0001
Gender	F	1.5831	1.4658	25	1.08	0.2904
Gender	M	0
Age		0.8268	0.07911	25	10.45	<.0001
Age*Gender	F	-0.3504	0.1239	25	-2.83	0.0091
Age*Gender	M	0

Type 3 Tests of Fixed Effects

	Num	Den		
Effect	DF	DF	F Value	Pr > F
Gender	1	25	1.17	0.2904
Age	1	25	110.54	<.0001
Age*Gender	1	25	7.99	0.0091

Estimates

		Standard			
Label	Estimate	Error	DF	t Value	Pr > t

gap at 14	-3.3231	0.8403	25	-3.95	0.0006
-----------	---------	--------	----	-------	--------

It is very interesting to compare the estimated within-subject variance matrices (and the corresponding correlation matrices) for the two models above. For the univariate design, the variance matrix has equal diagonal elements (the variance is assumed to be the same on each occasion) and equal off-diagonal elements (all pairwise covariances – and correlations – are the same). In the MANOVA model, the variances are estimated separately as are the covariances. All pairwise correlations are free to be different. Neither extreme seems

desirable in this case.

8.3.3 Random Intercept Model with Autocorrelation

We now use a mixed model with random intercept and autocorrelation. This will create a model that is intermediate between the two models above. It illustrates the use of both the RANDOM, for \mathbf{T} , and the REPEATED, for Σ , statements in the same model.

```
proc mixed data=pr method=ml covtest asycov asycorr;
  class Person Gender;
  model y = Gender Age Gender*Age / s;
  random INT / type=fa0(1) subject=Person
    g v vcorr;
  repeated / type = AR(1) subject=Person r rcorr;
  estimate 'gap at 14' Gender 1 -1 Age*Gender 14 -14;
run;
```

Partial output:

Estimated R Matrix for Person 1				
Row	Col1	Col2	Col3	Col4
1	1.8174	-0.1179	0.007655	-0.00050
2	-0.1179	1.8174	-0.1179	0.007655
3	0.007655	-0.1179	1.8174	-0.1179
4	-0.00050	0.007655	-0.1179	1.8174

1. This is $\hat{\Sigma}$. SAS tells us that this $\hat{\Sigma}$ for 'Person 1' but we have made no provisions to fit different $\hat{\Sigma}$ s for different people so this is $\hat{\Sigma}$ for everyone. We can fit different $\hat{\Sigma}$ s for boys and girls by using the GROUP = Gender option in the REPEATED statement. We can also use the GROUP option in the RANDOM statement to fit different values of \mathbf{T} for the two genders.

2. Note that the estimated autocorrelation between consecutive observations is **negative**: -0.1179, probably the opposite of what we expected. This could be because the random intercept is already doing a good enough job of accounting for correlation through $\mathbf{Z}\hat{\mathbf{T}}\mathbf{Z}'$. But if one has a close look at the data, it could also be the work of an anomalous observation that exhibits what looks like a wild zig-zag. It is left as an exercise to re-analyze the data without this observation.

Estimated R Correlation Matrix for Person 1				
Row	Col1	Col2	Col3	Col4
1	1.0000	-0.06490	0.004212	-0.00027
2	-0.06490	1.0000	-0.06490	0.004212
3	0.004212	-0.06490	1.0000	-0.06490
4	-0.00027	0.004212	-0.06490	1.0000

Estimated G Matrix			
Row	Effect	Person	Col1
1	Intercept	1	3.0904

Estimated V Matrix for Person 1				
Row	Col1	Col2	Col3	Col4
1	4.9078	2.9724	3.0980	3.0899
2	2.9724	4.9078	2.9724	3.0980
3	3.0980	2.9724	4.9078	2.9724
4	3.0899	3.0980	2.9724	4.9078

This is $\hat{\mathbf{V}}_1 = \hat{\mathbf{\Sigma}} + \mathbf{Z}_1\hat{\mathbf{T}}\mathbf{Z}'_1$ which does not vary under this model from person to person.

Estimated V Correlation Matrix for Person 1				
Row	Col1	Col2	Col3	Col4
1	1.0000	0.6057	0.6312	0.6296
2	0.6057	1.0000	0.6057	0.6312
3	0.6312	0.6057	1.0000	0.6057
4	0.6296	0.6312	0.6057	1.0000

Covariance Parameter Estimates	
	Standard Z

Cov Parm	Subject	Estimate	Error	Value	Pr > Z
FA(1,1)	Person	1.7579	0.2744	6.41	<.0001
AR(1)	Person	-0.06490	0.1612	-0.40	0.6872
Residual		1.8174	0.3078	5.91	<.0001

We see that the AR(1) parameter is not significantly different from 0.

Asymptotic Correlation Matrix of Estimates				
Row	Cov Parm	CovP1	CovP2	CovP3
1	FA(1,1)	1.0000	-0.1410	-0.1147
2	AR(1)	-0.1410	1.0000	0.3728
3	Residual	-0.1147	0.3728	1.0000

This output (produced by the ASYCORR option) shows the correlation between estimated covariance parameters. It is a good idea to consider this matrix to assess *collinearity* between covariance parameter estimates. Here, we see that there is a negative correlation between the AR(1) parameter and the FA(1,1) parameter that is equal to $\sqrt{\tau_{00}}$. We expected this because the correlation can be explained by one parameter or the other. It is almost surprising that the correlation is so low.

Fit Statistics	
-2 Log Likelihood	428.5
AIC (smaller is better)	442.5
AICC (smaller is better)	443.6
BIC (smaller is better)	451.6

Null Model Likelihood Ratio Test		
DF	Chi-Square	Pr > ChiSq
2	49.76	<.0001

This is a test of the hypothesis that $\mathbf{T} = \mathbf{0}$ and that the autocorrelation is 0, i.e. Σ has the usual scalar form $\sigma^2\mathbf{I}$.

Solution for Fixed Effects

Effect	Gender	Estimate	Standard Error	DF	t Value	Pr > t
Intercept		16.3140	0.9388	25	17.38	<.0001
Gender	F	1.0648	1.4709	79	0.72	0.4713
Gender	M	0
Age		0.7862	0.07400	79	10.62	<.0001
Age*Gender	F	-0.3072	0.1159	79	-2.65	0.0097
Age*Gender	M	0

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Gender	1	79	0.52	0.4713
Age	1	79	119.11	<.0001
Age*Gender	1	79	7.02	0.0097

Label	Estimate	Standard Error	DF	t Value	Pr > t
gap at 14	-3.2358	0.8113	79	-3.99	0.0001

8.3.4 Comparing Different Covariance Models

SEE HANDOUT

8.3.5 Exercises on Pothoff and Roy

1. Redo the random intercept with autocorrelation analysis without the apparent anomalous observation mentioned above. Does the estimate autocorrelation change as expected? Is there evidence that an autocorrelation term is needed?
2. Redo the analysis with a random slope model with and without autocorrelation. Compare the evidence for autocorrelation in the context of this model with the previous model.
3. Is there evidence of heteroscedasticity between boys and girls? (Use the GROUP option in the REPEATED or in the RANDOM statement depending on the way in which you

are choosing to model heteroscedasticity. Note that you can compare **nested models** differ only in their covariance structure (i.e. the same MODEL statement) by taking the difference of their deviances (-2 Log Likelihood) and comparing this difference to a Chi-Squared distribution with degrees of freedom equal to the difference in the number of parameters.

An easier way to do this in SAS is to compare the values shown for ‘Null Model Likelihood Ratio Test’ using the difference in degrees of freedom as the degrees of freedom for the Chi Square.

4. Consider a random intercept and slope model (perhaps with heteroscedasticity by Gender). Is there evidence that $\tau_{01} \neq 0$? In this particular situation would this be a sensible hypothesis to consider?

8.4 Panel Study of Income Dynamics (PSID)

The data used from the PSID comes from Cornwell and Rupert (1988):

The data for our analysis are drawn from years 1976–1982 of the non-Survey of Economic Opportunity portion of the Panel Study of Income Dynamics (PSID). The individuals in our sample are 595 heads of household between the ages of 18 and 65 in 1976, who report a positive wage insome private, non-farm employment for all 7 years. So, for each individual, we have education (ED), years of full-time work experience (EXP), weeks worked (WKS), occupation (OCC = 1, if the individual has blue-collar occupation), industry (IND = 1, if the individual works in a manufacturing industry), residence (SOUTH = 1, SMSA = 1 if the individual resides in the south, or in a stadard metropolitan statistical area), marital status (MS = 1, if the individual is mattied), union coverage (UNION = 1, if the indivudal’s wage is set by a union contract), sex and race (FEM = 1, BLK = 1, if the individual is female or black).

The following analysis implements in PROC MIXED an analysis that parallels the instrumental variable estimator proposed by Hausman and W. (1981). The model has the form

$$Y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_i\boldsymbol{\gamma} + \alpha_i + \varepsilon_{it}, \quad i = 1, \dots, N; \quad t = 1, \dots, T \quad (158)$$

where \mathbf{x}_{it} is a $K \times 1$ vector of time-varying explanatory variables and \mathbf{z}_i is a $G \times 1$ vector of time-invariant explanatory variables. We assume $\alpha_i \sim N(0, \sigma_\alpha^2)$ and $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$. Letting \mathbf{V} represent a matrix of dummy variables for the individuals, we can write:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{V}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \quad (159)$$

with $\boldsymbol{\alpha} \sim \mathbf{N}(\mathbf{0}, \sigma_{\alpha}^2 \mathbf{I})$ and, independently, $\boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0}, \sigma_{\varepsilon}^2 \mathbf{I})$. The $\boldsymbol{\alpha}$ vector might be correlated with parts of \mathbf{X} and \mathbf{Z} . This approach further classifies predictors into 4 groups depending on whether they are assumed to be correlated with $\boldsymbol{\alpha}$:

	Uncorrelated with effects	Correlated with effects
Time variant	\mathbf{X}_1	\mathbf{X}_2
Time invariant	\mathbf{Z}_1	\mathbf{Z}_2

The estimator uses an instrument set:

$$\mathbf{A} = (\mathbf{Q}_{\mathbf{V}}\mathbf{X}_1, \mathbf{Q}_{\mathbf{V}}\mathbf{X}_2, \mathbf{P}_{\mathbf{V}}\mathbf{X}_1, \mathbf{Z}_1) \quad (160)$$

where $\mathbf{P}_{\mathbf{V}}$ and $\mathbf{Q}_{\mathbf{V}}$ are the projection matrices into the space spanned by \mathbf{V} and its orthogonal complement, respectively. In simple words, $\mathbf{P}_{\mathbf{V}}$ takes the individual subject means and $\mathbf{Q}_{\mathbf{V}}$, the residual from the subject means.

Using an instrument set is equivalent to controlling for the residuals that you get when you predict the predictors with the instruments. (You've got to stop and think about it). Now, it's easy to see that \mathbf{X}_1 and \mathbf{Z}_1 are perfectly predicted by \mathbf{A} , so they produce no residuals. For \mathbf{X}_2 and \mathbf{Z}_2 this involves first taking individual means for \mathbf{X}_2 , i.e. $\mathbf{P}_{\mathbf{V}}\mathbf{X}_2$, and then finding the residuals of $\mathbf{P}_{\mathbf{V}}\mathbf{X}_2$ and \mathbf{Z}_2 on the space $\mathbf{A}_{\mathbf{V}} = (\mathbf{P}_{\mathbf{V}}\mathbf{X}_1, \mathbf{Z}_1)$.

Cornwell and Rupert (1988) classify the PSID variables as follows:

	Uncorrelated with effects	Correlated with effects
Time variant	WKS, SOUTH, SMSA, MS	EXP, EXP ² , OCC, IND, UNION
Time invariant	FEM, BLK	ED

The following code illustrates a first step in this procedure in controlling for individual averages of time-varying predictors that are suspected of being correlated with effects. Although the coefficients of time-invariant predictors would not be adjusted as they would be with the Hausman-Taylor approach, the space of time-invariant predictors will be the same so that the coefficients of time-varying predictors should be similarly adjusted.

```
data psid;
  set psid.psid;
  run;

proc contents data = psid;
  run;

proc means data = psid print;
```

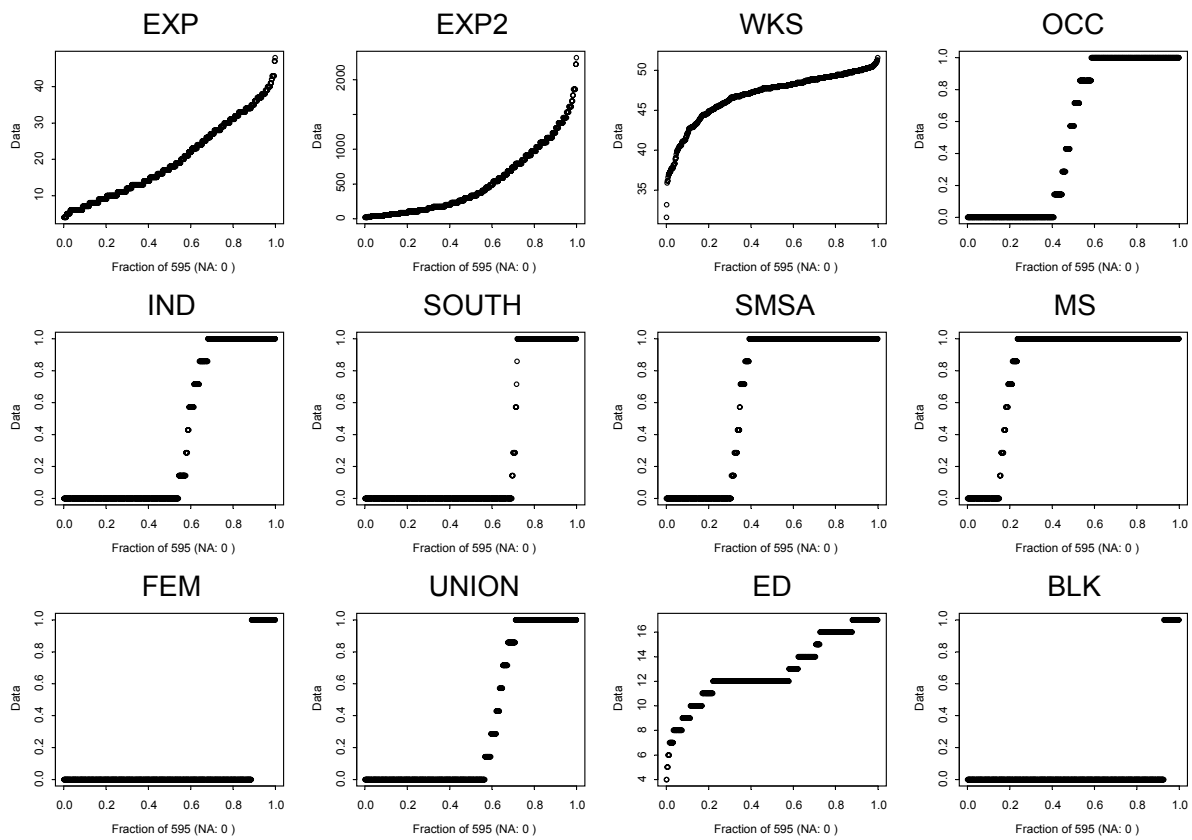


Figure 3: Quantile plots of individual means of some PSID variables.

```

run;

proc sort data = psid;
  by i;
run;

/* we can adjust for inflation */
data psid;
  set psid;
  if T = 1 then cpi = 56.9;
  else if T = 2 then cpi = 60.6;
  else if T = 3 then cpi = 65.2;
  else if T = 4 then cpi = 72.6;
  else if T = 5 then cpi = 82.4;
  else if T = 6 then cpi = 90.9;
  else if T = 7 then cpi = 96.5;
/* CPI adjusted log wage
   multiplying by 100
   turns coefficients into percentages */
  alwage = 100*(lwage - log(cpi/100)); /
run;

proc means data = psid noprint;
  var exp occ ind union ;
  by i;
  output out = agg mean = expm occm indm unionm;
run;

proc sort data = agg;
  by i;
run;

data psid;
  merge psid agg;
  by i;
  expm2 = expm*expm;
run;

```

```

/* model not adjusting for contextual effects */
proc mixed data = psid covtest cl asycov asycorr;
  model alwage = exp exp2 wks occ ind union
    south smsa ms ed fem blk /
    s corrb covb covbi cl
    ddfm = satterth;
  random int exp / sub = i g gc type = fa0(2);
run;

```

```

/* model adjusting for contextual effects */
proc mixed data = psid covtest cl asycov asycorr;
  model alwage = exp exp2 wks occ ind union
    south smsa ms ed fem blk
    expm expm2 occm indm unionm/
    s corrb covb covbi cl
    ddfm = satterth;
  random int exp / sub = i g gc type = fa0(2);
run;

```

```

/* Compare the coefficients in these two models */
/* Explore larger random models */

```

8.5 Non-linear Growth Curves with NLMIXED

SEE HANDOUTS

8.6 Logistic Mixed Regression with NLMIXED

SEE HANDOUTS

9 Bibliography

References

- Ahn, C. H. (1990). Diagnostics for heteroscedasticity in mixed linear models (stma V32 2489). *Journal of the Korean Statistical Society* 19, 171–175.
- Albert, J. and S. Chib (1997). Bayesian tests and model diagnostics in conditionally independent hierarchical models. *Journal of the American Statistical Association* 92, 916–925.
- Banerjee, M. and E. W. Frees (1997). Influence diagnostics for linear longitudinal models. *Journal of the American Statistical Association* 92, 999–1005.
- Beckman, R. J., C. J. Nachtsheim, and R. D. Cook (1987). Diagnostics for mixed-model analysis of variance (corr: V32 p241). *Technometrics* 29, 413–426.
- Bijleveld, C. C. and L. J. T. Van der Kamp (1998). *Longitudinal Data Analysis: Designs, Models and Methods*. London: Sage.
- Bradlow, E. T. and A. M. Zaslavsky (1997). Case influence analysis in Bayesian inference. *Journal of Computational and Graphical Statistics* 6, 314–331.
- Bryk, A. S. and S. W. Raudenbush (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage, Newbury Park.
- Christensen, R. and E. J. Bedrick (1997). Testing the independence assumption in linear models. *Journal of the American Statistical Association* 92, 1006–1016.
- Christensen, R., L. M. Pearson, and W. Johnson (1992). Case-deletion diagnostics for mixed models. *Technometrics* 34, 38–45.
- Cornwell, C. and P. Rupert (1988). Efficient estimation with panel data: An empirical comparison of instrumental variables estimators. *Journal of Applied Econometrics* 3, 149–155.
- Cox, D. R. and N. Wermuth (1996). *Multivariate Dependencies – Models, analysis and interpretation*. Chapman & Hall, London.
- Davidian, M. and D. M. Giltinan (1995). *Nonlinear Models for Repeated Measurement Data*. Chapman and Hall, Boca Raton.
- Diggle, P. J., K.-Y. Liang, and S. L. Zeger (1994). *Analysis of Longitudinal Data*. Clarendon Press.
- Draper, N. R. and H. Smith (1981). *Applied Regression Analysis* (2nd ed.). Wiley Series in Probability and Mathematical Statistics. Wiley, New York.

- Efron, B. and C. Morris (1977). Stein's paradox in statistics. *Scientific American* 236, 119–127.
- Fox, J. (1991). *Regression Diagnostics*. Sage.
- Hausman, J. A. and T. W. (1981). Panel data and unobservable individual effects. *Econometrica* 49, 1377–1399.
- Helen Brown, R. P. (2000). *Applied Mixed Models in Medicine: Statistics in Practice*. N. Y.: Wiley.
- Ho, Y.-Y., M. Peruggia, and T. J. Santner (1995). Diagnostics for hierarchical Bayesian repeated measures models. In *Computing Science and Statistics. Statistics and Manufacturing with Subthemes in Environmental Statistics, Graphics and Imaging. Proceedings of the 27th Symposium on the Interface*, pp. 387–391. Interface Foundation of North America (Fairfax Station, VA).
- Hocking, R. R. and R. H. Bremer (1987). Estimation of variance components in mixed factorial models including model-based diagnostics. In *Proceedings of the SAS Users Group International Conference*, Volume 12, pp. 1162–1167. SAS Institute (Cary, NC).
- Hocking, R. R., J. W. Green, and R. H. Bremer (1989). Variance-component estimation with model-based diagnostics. *Technometrics* 31, 227–239.
- Hodges, J. S. (1998). Some algebra and geometry for hierarchical models, applied to diagnostics (disc: P521-536). *Journal of the Royal Statistical Society, Series B, Methodological* 60, 497–521.
- Johnston, J. and J. DiNardo (1997). *Econometric Methods (Fourth Edition)*. McGraw Hill International Editions.
- Kreft, I. G. G. and J. de Leeuw (1998). *Introducing Multilevel Modeling*. Introducing Statistical Methods. Sage Publications.
- Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics* 38, 963–974.
- Littell, R. C., G. A. Milliken, W. W. Stroup, and R. D. Wolfinger (1996). *SAS System for Mixed Models*. SAS Institute, Cary, N.C.
- McCulloch, C. E. and S. R. Searle (2000). *Generalized, Linear, and Mixed Models*. Wiley Series in Probability and Statistics. Texts, References, and Pocketbooks Section. N. Y.: Wiley.
- Monette, G. (1990). Geometry of multiple regression and interactive 3-d graphics. In J. Fox and J. S. Long (Eds.), *Modern Methods of Data Analysis*, pp. 209–256. Sage (Newbury Park, CA; London).

- Pinheiro, J. C. and D. M. Bates (2000). *Mixed-Effects Models in S and S-PLUS*. Statistics and Computing. Springer.
- Potthoff, R. F. and S. N. Roy (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika* 51, 313–326. Opt.
- SAS Institute Inc. (2000a). *The MIXED Procedure*, Chapter 37, pp. 1947–2088. SAS, (Cary, N.C.).
- SAS Institute Inc. (2000b). *SAS OnlineDoc: Version 7–1*. SAS, (Cary, N.C.).
- Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, 323–355.
- Snijders, T. A. B. and R. J. Bosker (1999). *Multilevel Analysis: An introduction to basic and advanced multilevel modeling*. Sage, London.
- Venables, W. V. and B. D. Ripley (1999). *Modern Applied Statistics with S-Plus (3rd ed.)*. Springer, New York.
- Verbeke, G. and G. Molenberghs (Eds.) (1997). *Linear Mixed Models in Practice: A SAS-Oriented Approach*. Springer, New York.
- Vonesh, E. F. and V. M. Chinchilli (Eds.) (1997). *Linear and nonlinear models for the analysis of repeated measurements*. Dekker, New York.
- Wernecke, K.-D., G. Kalb, and E. Stürzebecher (1988). On classification strategies in medical diagnostics (with special reference to mixed models). In *Classification and Related Methods of Data Analysis*, pp. 299–306. North-Holland/Elsevier (Amsterdam; New York).

A Synopsis of SAS commands in PROC MIXED

For a complete list of SAS commands for PROC MIXED, see SAS on-line documentation SAS Institute Inc. (2000b). The following is a list of the key commands and options.

The concepts behind some of these commands have not been covered in the material so far and will be seen later in the course.⁴

PROC MIXED : The following are some of the commonly used options on the PROC statement:⁵

Input: DATA = dsname

Output:

ASYCOV ASYCORR : asymptotic covariance/correlation of the variance/covariance (\mathbf{T} and Σ) parameter estimates.

CL : confidence limits for above.

COVTEST : tests for above.

RATIO : produces \mathbf{T}/σ^2 .

LOGNOTE : for very long runs – so SAS will show signs of life.

ITDETAILS : for when things seem to go wrong.

MMEQ : matrices of coefficients of “mixed model equations” which are extensions of the OLS “normal” equations.

Method:

METHOD = [REML] | ML : restricted or full likelihood (choose ML if you will compare log-likelihoods between models with different fixed models [i.e. different MODEL statements] but ML sometimes has trouble converging with near singular \mathbf{T} [G] matrices).

EMPIRICAL : uses robust “sandwich” estimator to estimate variance-covariance of fixed effect estimates. The “sandwich” estimate uses the observed subject-to-subject variance instead of using the inverse Fisher information of the normal model.

Algorithm:

CONVF | CONVG | [CONVH] : choose convergence criterion.

⁴SAS documentation uses \mathbf{G} instead of \mathbf{T} , \mathbf{R} instead of Σ , β instead of γ , and γ instead of \mathbf{u} . \mathbf{Y} , \mathbf{X} and \mathbf{Z} are used the same way in SAS and in Snijders and Bosker (1999) and Bryk and Raudenbush (1992)! This multiplicity of dialects is one consequence of the multi-cultural history of mixed models.

⁵Default choices are shown in [square brackets]. Alternative choices are separated by vertical bars | .

MAXFUNC=[150] MAXITER=[50] : maximum number of function evaluations at each iteration and maximum number of iterations.

NOPROFILE : treats σ^2 like other parameters. e.g. so it can be “held” with the HOLD option of the PARMS statement.

CLASS vars : names categorical variables used in analysis.

MODEL y = x1 x2 x1*x2 : specify fixed model (intercept is included by default)

Output:

S : show “solution,” i.e. estimated values of parameters.

CORRB COVB COVBI : variances/covariances/correlations and inverse covariances of estimates.

CL : confidence intervals.

OUTPREDM = ds1 OUTPRED = ds2 : output datasets with ‘population’ predicted values (no EBLUPS) and ‘cluster-wise’ predicted values (including contribution of EBLUPS).

Method:

DDFM = SATTERTH : uses Satterthwaite approximation for denominator degrees of freedom. A newer method KENWARDROGER might be better.

NOINT : no intercept.

RANDOM INT x1 x2 : specify random model.

Output:

G GC GCORR GI : print the \mathbf{T} matrix, its Cholesky factor (useful to determine rank), its covariances as correlations, the inverse of \mathbf{T} . Note that the Cholesky factor can be useful to detect rank.

S : EBLUPS. Beware if many clusters.

V VCORR : within-cluster variance.

Method:

SUB = var : variable identifying cluster. If omitted analysis is done as if for a single cluster. Note that the data set should be sorted on this variable.

GRP = var : variable identifying groups with possibly different \mathbf{T} matrices.

TYPE = [VC] | UN | FAO(q) : There are many more variance-covariance structures but these three are generally suitable for random effects. **VC** is appropriate for homoscedastic categorical variables. **UN** and **FAO(q)** where q is the order of \mathbf{T} both generate variance matrices with no constraints except that matrices fit with **UN** need not be non-negative definite. Some parameters in both forms can be constrained with the **PARMS** statement.

REPEATED : specifies within cluster variance structure. It is especially useful to model autocorrelation for longitudinal data.

Output:

R RC RCI RCORR RI : various reports on within cluster variance.

Method:

SUB = var : same as **RANDOM** statement.

GRP = var : variable identifying groups with potentially different values of Σ .

LOCAL = POM(data set) | EXP(var) : allows variance to vary as a power of $E(Y)$ or as a function of predictors.

TYPE = ANTE(1) | AR(1) | ARH(1) | ARMA(1,1) | UN | SP(POW)(var) : default is $\sigma^2\mathbf{I}$. Other options are discussed later in notes. They are used primarily for longitudinal data or to fit multilevel multivariate models.

ESTIMATE CONTRAST LSMEANS : Used to estimate $E(Y|...)$ or differences. See SAS Institute Inc. (2000b) and notes.

PARMS : sets initial values for \mathbf{T} and Σ parameters.

Method:

HOLD = i, j : keeps parameters in position i, j fixed at initial value.

A.1 PROC MIXED all-dressed

This is an example of calling **PROC MIXED** with many options. The model used has random effects for two inner variables: **SES** and **FEMALE**. A constrained Cholesky parametrization is used to fit a covariance of 0 between the random effects for **SES** and **FEMALE**.

```

PROC MIXED DATA = hs ASYCOV ASYCORR CL COVTEST;
  CLASS SCHOOL;
  MODEL mathach = SES SECTOR FEMALE SES*SECTOR /
    S CORRB COVB COVBI CL
    DDFM = SATTERTH
    OUTPREDM = hsm OUTPRED = hsc ;
  RANDOM SES FEMALE INT / /* note: INT last */
  SUB = SCHOOL
  TYPE = FA0(3)
  S
  G GC GCORR GI
  V VCORR; /* big and not very interesting here */
  ESTIMATE 'Cath-Pub | low SES'
    SECTOR 1 SES*SECTOR -2 / CL E ;
  ESTIMATE 'Cath-Pub | high SES'
    SECTOR 1 SES*SECTOR +2 / CL E ;
  PARS (1) (0) (1) (1) (1) (1) / HOLD = 2;
RUN;

```

A.2 Exercises

Apply the models above to the “hs” data set which is a subsample of 40 schools from the data set reported in Bryk and Raudenbush (1992).

In each case use the ESTIMATE statement to test plausible questions. Sketch the fitted response lines.

One-way ANOVA

We use the HS data set and fit five models for comparison.

1. single mean for all the data
2. school effect coded with “true contrasts”, i.e. each column of the coding matrix sums to 0. We will use helmert contrasts but, for our purposes, it doesn’t matter what kind of contrast is used.

3. school effect coded with “sample size” contrasts, i.e. the columns of the contrast matrix are orthogonal to sample size.
4. school effect coded with ...
5. lme

Tricks

Simplifying the random model

In many applications, there are many inner variables and no a priori reason to exclude some from the random model. Also, there might not be any compelling reasons to set some covariances to zero in order to reduce the variance component.

The variance of estimators is determined by the design and it is a relatively simple matter to re-express predictor variables to achieve uncorrelated estimators, for example. The variance components (variance-covariance of random effects) does not necessarily bear any relationship with the design. In this section we consider the problem of re-expressing predictors associated with random effects in order to simplify the random model.

Let the model, in Laird-Ware (1982) form be:

$$Y_j = X_j\beta + Z_jb_j + \epsilon_j \tag{161}$$

Suppose $b_j \sim N_q(0, D)$ and $\epsilon_j \sim N_{n_j}(0, \Sigma_j)$. Frequently, the estimated D will be nearly singular and the model can be reduced by transforming Z_j to Z_jA where A is $q \times r$ with $r < q$.

We want to keep the random portion of the model, i.e.:

$$Z_jb_j = (Z_jA)b_j^* \tag{162}$$

If $b_j^* \sim N(0, I)$ then $AA' = D$.

Essentially,

If D has spectral decomposition: $D = \Gamma_r\Lambda_r\Gamma_r'$ with Γ_r of size $q \times r$ and Λ_r an $r \times r$ diagonal matrix with positive diagonal elements, the general form of a matrix with $AA' = D$ is $\Gamma_r\Lambda_r^{1/2}Q'$ where $Q'Q = I$.

=====

From Longord (1993)

If we transform Z_j but post-multiplying by a matrix A , we pre-multiply b_j by A^{-1} and the resulting variance of b_j is $A^{-1}DA^{-1}$.

Assume the first column of Z_j consists of 1s. For the transformation, A , to preserve the intercept, it will have a first column $(1, 0, \dots, 0)'$ Now consider the general effect

=====

Good idea: develop a program that plots the data ellipse for two variables and the “variation pattern” ellipse centered at the point of minimum variance of Y and with ellipse showing variance isoquants.

If the centres and principal axes of these two ellipses coincide, we can “orthogonalize” both the fixed and random portions of the model with a single re-expression.

=====

Fitting GLMs:

Don't forget: `nlmeControl (sigma = 1)`. Note that this ends up being equivalent to independence of multinomials in a random intercept poisson model. Note that holding `sigma=1` might provide more info on variance pattern – although maybe this isn't a problem if the algo uses the relative precision matrix !?!

Deal with convergence problems.... somehow.

Understanding “multicollinearity of parameters” This is especially difficult between fixed and random, and among random parameters. A first step is provided by plotting the variance pattern ellipse [see if this is in accord with idea of a pattern matrix in factor analysis].

Where else can the parameter variance / variance isoquant duality be exploited?

Distinguish between marginal likelihood and profile likelihood. What are the differences in applications? How about conditional likelihoods? Perhaps we can have a field guide to likelihoods in mixed models.

Panel data in econometrics and multilevel models

It is useful to comment on the similarities between the material covered here and some issues in the analysis of panel data in econometrics as presented in Johnston and DiNardo (1997).

Questions and Issues

1. The random model:
 - (a) effect on convergence – balancing sds of effects
 - (b) finding a simpler structure: factoring VarCorr matrix? Are there sensible simplifications that make sense in the context of the model?

- (c) interpreting the `ms` parameters: are they negative of the log-diagonal parametrization of the *relative precision* matrix Δ .
- (d) Consider Cox and Wermuth's questions re parsimony. Do we contribute to understanding when we find a parsimonious variance-covariance matrix? Is it important to try?

2. Using factors:

- (a) How to get desired definition for a parameter?

3. Clarify role of `aov`, `varcomp` and `lme`.

4. Predictor centering: what it does, effect on interpretation, effect on interpretation of $\text{Var}(b)$.

Should you center the design or the variance?

Orthogonalize the design of the variance?

5. TOUGH ISSUE: Reducing variance parameters.

6. INVARIANCE: when something matters and it shouldn't. **need for intercept-slope covariances**

7. effect on shrinking

8. p. 72 of SB

no

B TODO

1. Consider use of RANDOM model with different variables than fixed model.

2. Consider rank deficient FAs to choose between collinear optional vars in RANDOM model, e.g. different centerings.

3. Work on PSID and implementing instrumental variables [see pdf paper]

4. Add choleski trick

5. Add recentering trick
6. Hierarchical R^2
7. Add splines
8. Add periodic functions

Handouts:

PROC MIXED,
PROC NL MIXED,
MLwiN ?
STATA ?

Questions to address:

2SLS
Conditional logit

Orphaned topics:

R^2

Outline

General:

Start with Bryk and Raudenbush: basic concepts

Talk about 1 school, slope and intercept

Talk about 2 schools what you can say

Interaction, relocation. Paradox: relocate changes school effect

Use ellipses and "lines through center". Sig main with inter
but not without.

Discuss issue of marginality: 3 levels of understanding

Talk about all schools. How could we answer a simple question: Where should a low SES person send their child, Cath, Public, Low or high SES? At some point: CAUTION in interpretation but leave to later to make what follows more interesting.

How can we compare Cath vs. Public: [good opportunity to talk about problems]

Aggregate? (problems: Simpson or Robinson)

Special contrast? (target of inference)

2 stage: not bad but what to do with variance? OK with perfect balance
[use Mahalanobis thing to show equality if n , σ^2 are same]
Discuss classical nested ANOVA to situate.

Problems without mixed

In a sense, mixed models allow estimation in unbalanced ...

Add detailed stuff from handout

Somewhere around here: Between vs. Within (with ellipses: first use consistent ellipses that lineup and then use inconsistent ellipses)

How GLS combines optimally.

Mixed models:

Combining information between and within: how? shrinkage
(use details in hs2)

Combined model: $Y = Xb + Zb + e$

Implementation in SAS

Using SAS: interpretation of output

Emphasize diagnostics

Strategy: Start small, work up

Variance of predictors:

patterns: full
less than full
caution re marginality

Mention examples: families, regions,

When to use fixed, when to use mixed. Warn about two ends of continuum:

Fixed (no groups) - mixed - fixed groups
fixed or mixed is issue at both ends ... sometimes confusing!

Longitudinal:

Concepts to present:

What is it: Data collected over time: Use Potthoff and Roy for example.

Traditional: Repeated measures: ANOVA and MANOVA approach

Problems:

- two extremes:

ANOVA: minimal variance parameters: assumes compound symmetry, equal variances (not painful), equal correlation between ANY pair (painful over time).

MANOVA: maximal parameters - everything goes -nothing

in between

- with MANOVA, can't gain efficiency in variance estimation from smaller means model

- Need fully balanced data for MANOVA, ANOVA not efficient with less than fully efficient.

What's missing: ability to deal with richer growth models, covariance patterns, that take time order into account (i.e. observations close in time more correlated than observations far in time)

Models over time: growth curves: linear, quadratic, more general, opportunity to be creative, interrupted,

Two examples: Linear using transformation (use IQ with predetermined transformation)

non-linear: use IQ with profile likelihood non-normal response: Build on mixed (allow non-diagonal patterns for Sigma)

IQ

Non-linear: migraines

Survey: we'll see

Caution: An exaggerated estimate of σ^2 will be problematic. Because more parameter variance will appear to have been explained by σ^2 than should be, and the estimates of T will be too small resulting in non-convergence, etc.

#

C A synopsis of topics for day 2:

1. Simpson's and Robinson's paradoxes: their role in mixed models.
2. Cluster averages and within-cluster residuals: paradox lost.
3. Structures for random-effects variance-covariance matrix.
4. Hypothesis testing for fixed and random effects.
5. What to try when the model doesn't converge – and even when it does: centering and balancing for mixed models.
6. Realistic model building: small to big or big to small.
7. Multilevel R^2 .
8. Diagnostics and exploratory model building in mixed models.
9. Some formal properties of mixed models: what models are invariant under what transformations?

D Changes

1. June 22, 2001: Fixed error in of omitting σ^2 in off-diagonal entries in equations 140, 139, and 137.
2. Nov. 22, 2001: Added negative sign to multivariate formula for X_{min} .