Sociology 761 John Fox

Lecture Notes

Introduction to Mixed-Effects Models for Hierarchical and Longitudinal Data

1. Introduction (and Review)

▶ The *standard linear model*

$$\begin{aligned} y_i &= \beta_1 + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i \\ \varepsilon_i &\sim \text{NID}(0, \sigma^2) \end{aligned}$$

assumes independently sampled observations, and hence independent errors ε_i.

- In matrix form,

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon} &\sim \mathbf{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n) \end{aligned}$$

  where
  - $\mathbf{y} = (y_1, y_2, ..., y_n)'$ is the response vector;
  - $\mathbf{X}$ is the model matrix, with typical row $\mathbf{x}_i' = (1, x_{2i}, ..., x_{pi})$;
  - $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)'$ is the vector of regression coefficients;
  - $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)'$ is the vector of errors;
  - $\mathbf{N}_n$ represents the n-variable multivariate-normal distribution;

– $\mathbf{0}$ is an $n \times 1$ vector of zeroes; and
– $\mathbf{I}_n$ is the order-n identity matrix.

- The standard linear model has one *random effect*, the error term ε_i, and one *variance component*, $\sigma^2 = \text{Var}(\varepsilon_i)$.
- When the assumptions of the standard linear model hold, *ordinary-least-squares* (*OLS*) regression provides maximum-likelihood estimates of the regression coefficients,

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- The MLE of the error variance σ^2 is

$$\widehat{\sigma}^2 = \frac{\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)}{n}$$

- $\widehat{\sigma}^2$ is a biased estimator of σ^2; usually, the unbiased estimator

$$s^2 = \frac{\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)}{n - p}$$

is preferred.

▶ The standard linear model and OLS regression are generally inappropriate for dependent observations.

- Dependent (or *clustered*) data arise in many contexts, the two most common of which are hierarchical data and longitudinal data.

- *Hierarchical data* are collected when sampling takes place at two or more levels, one *nested* within the other. Some examples:
  - Students within schools (two levels).
  - Students within classrooms within schools (three levels).
  - Individuals within nations (two levels).
  - Individuals within communities within nations (three levels).
  - Patients within physicians (two levels).
  - Patients within physicians within hospitals (three levels).
- There can also be non-nested multi-level data — for example, high-school students who each have multiple teachers.

- *Longitudinal data* are collected when individuals (or other units of observation) are followed over time. Some examples:
  - Annual data on vocabulary growth among children.
  - Biannual data on weight-preoccupation and exercise among adolescent girls.
  - Data collected at irregular intervals on recovery of IQ among coma patients.
  - Annual data on employment and income for a sample of adult Canadians.
- In all of these cases, it is not generally reasonable to assume that observations within the same higher-level unit, or longitudinal observations within the same individual, are independent of one-another.

- *Mixed-effect models* make it possible to take account of dependencies in hierarchical, longitudinal, and other dependent data.
  - Unlike the standard linear model, mixed-effect models include more than one source of random variation — i.e., more than one random effect.
  - Mixed-effects models have been developed in a variety of disciplines, with varying names and terminology: *random-effects models* (statistics, econometrics), *variance and covariance-component models* (statistics), *hierarchical linear models* (education), *multi-level models* (sociology), *contextual-effects models* (sociology), *random-coefficient models* (econometrics), *repeated-measures models* (statistics, psychology).
  - Mixed-effects models have a long history, dating to Fisher and Yates's work on "split-plot" agricultural experiments.

  - What distinguishes modern mixed models from their predecessors is generality: for example, the ability to accommodate irregular and missing observations.

Sociology 761
Copyright © 2014 by John Fox

- Principal sources for these lectures on mixed models:
  - J. Fox, *Applied Regression Analysis and Generalized Linear Models, Third Edition*, Sage, 2002, Chapters 23 and 24.
  - J. Fox and S. Weisberg, *An R Companion to Applied Regression, Second Edition*, Sage, 2011, “Mixed-Effects Models in R” (Appendix, draft).
- Topics:
  - The linear mixed-effects model.
  - Modeling hierarchical data.
  - Modeling longitudinal data.
  - Generalized linear mixed models (time permitting).

2. The Linear Mixed-Effects Model

- This section introduces a very general linear mixed model, which we will adapt to particular circumstances.
- The *Laird-Ware form* of the linear mixed model:

$$\begin{aligned} y_{ij} &= \beta_1 + \beta_2 x_{2ij} + \cdots + \beta_p x_{pij} + b_{1i} z_{1ij} + \cdots + b_{qi} z_{qij} + \varepsilon_{ij} \\ b_{ki} &\sim N(0, \psi_k^2), \text{Cov}(b_{ki}, b_{k'i}) = \psi_{kk'} \\ & \quad b_{ki}, b_{k'i'} \text{ are independent for } i \neq i' \\ \varepsilon_{ij} &\sim N(0, \sigma^2 \lambda_{ijj}), \text{Cov}(\varepsilon_{ij}, \varepsilon_{ij'}) = \sigma^2 \lambda_{ijj'} \\ & \quad \varepsilon_{ij}, \varepsilon_{i'j'} \text{ are independent for } i \neq i' \end{aligned}$$

  where
  - y_{ij} is the value of the response variable for the jth of n_i observations in the ith of M groups or clusters.
  - $\beta_1, \beta_2, \ldots, \beta_p$ are the fixed-effect coefficients, which are identical for all groups.

- $x_{2ij}, \ldots, x_{pij}$ are the fixed-effect regressors for observation j in group i; there is also implicitly a constant regressor, $x_{1ij} = 1$.
- $b_{1i}, \ldots, b_{qi}$ are the random-effect coefficients for group i, assumed to be multivariately normally distributed, independent of the random effects of other groups. The random effects, therefore, vary by group.
  – The b_{ik} are thought of as random variables, not as parameters, and are similar in this respect to the errors ε_{ij}.
- $z_{1ij}, \ldots, z_{qij}$ are the random-effect regressors.
  – The zs are almost always a subset of the xs (and may include *all* of the x s).
  – When there is a random intercept term, $z_{1ij} = 1$.
- ψ_k^2 are the variances and $\psi_{kk'}$ the covariances among the random effects, assumed to be constant across groups.
  – In some applications, the ψs are parametrized in terms of a smaller number of fundamental parameters.

- ε_{ij} is the error for observation j in group i.
  – The errors for group i are assumed to be multivariately normally distributed, and independent of errors in other groups.
- $\sigma^2\lambda_{ijj'}$ are the covariances between errors in group i.
  – Generally, the $\lambda_{ijj'}$ are parametrized in terms of a few basic parameters, and their specific form depends upon context.
  – When observations are sampled independently within groups and are assumed to have constant error variance (as is typical in hierarchical models), $\lambda_{ijj} = 1$, $\lambda_{ijj'} = 0$ (for $j \neq j'$), and thus the only free parameter to estimate is the common error variance, σ^2.
  – If the observations in a “group” represent longitudinal data on a single individual, then the structure of the λs may be specified to capture serial (i.e., over-time) dependencies among the errors.

▶ The Laird-Ware model in matrix form:

$$
\begin{aligned}
\mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \\
\mathbf{b}_i &\sim \mathbf{N}_q(\mathbf{0}, \boldsymbol{\Psi}) \\
& \quad \mathbf{b}_i, \mathbf{b}_{i'} \text{ are independent for } i \neq i' \\
\boldsymbol{\varepsilon}_i &\sim \mathbf{N}_{n_i}(\mathbf{0}, \sigma^2\boldsymbol{\Lambda}_i) \\
& \quad \boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_{i'} \text{ are independent for } i \neq i'
\end{aligned}
$$

where

- $\mathbf{y}_i$ is the $n_i \times 1$ response vector for observations in the ith group.
- $\mathbf{X}_i$ is the $n_i \times p$ model matrix for the fixed effects for observations in group i.
- $\boldsymbol{\beta}$ is the $p \times 1$ vector of fixed-effect coefficients.
- $\mathbf{Z}_i$ is the $n_i \times q$ model matrix for the random effects for observations in group i.
- $\mathbf{b}_i$ is the $q \times 1$ vector of random-effect coefficients for group i.
- $\boldsymbol{\varepsilon}_i$ is the $n_i \times 1$ vector of errors for observations in group i.

- $\boldsymbol{\Psi}$ is the $q \times q$ covariance matrix for the random effects.
- $\sigma^2\boldsymbol{\Lambda}_i$ is the $n_i \times n_i$ covariance matrix for the errors in group i, and is $\sigma^2\mathbf{I}_{n_i}$ if the within-group errors are homoscedastic and independent.

▶ Notice that there are two things that distinguish the linear mixed model from the standard linear model:

(a) There are structured random effects $\mathbf{b}_i$ in addition to the errors $\boldsymbol{\varepsilon}_i$.

(b) The model can accommodate heteroscedasticity and dependencies among the errors.

3. Modeling Hierarchical Data

▶ Applications of mixed models to hierarchical data have become common in the social sciences, and nowhere more so than in research on education.

▶ I'll restrict myself to two-level models, but three or more levels can also be handled through an extension of the Laird-Ware model.

▶ The following example is borrowed from Raudenbush and Bryk, and has been used by others as well (though we will learn some things about the data that apparently haven't been noticed before).
- The data are from the 1982 "High School and Beyond" survey, and pertain to 7185 U.S. high-school students from 160 schools — about 45 students on average per school.
  - 70 of the high schools are Catholic schools and 90 are public schools.

- The object of the data analysis is to determine how students' math achievement scores are related to their family socioeconomic status.
  - We will entertain the possibility that the level of math achievement and the relationship between achievement and SES vary among schools.
  - If there is evidence of variation among schools, we will examine whether this variation is related to school characteristics — in particular, whether the school is a public school or a Catholic school and the average SES of students in the school.

▶ A good place to start is to examine the relationship between math achievement and SES separately for each school.
- 160 schools are too many to look at individually, so I sampled 20 Catholic school and 20 public schools at random.
- The scatterplots for the sampled schools are in Figures 1 and 2.

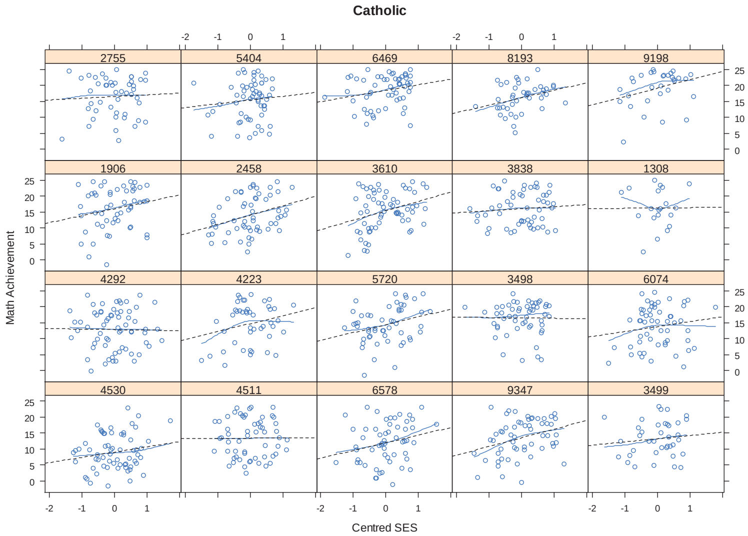

Figure 1. Math achievement by SES for students in 20 randomly selected Catholic schools. SES is centred at the mean of each school.

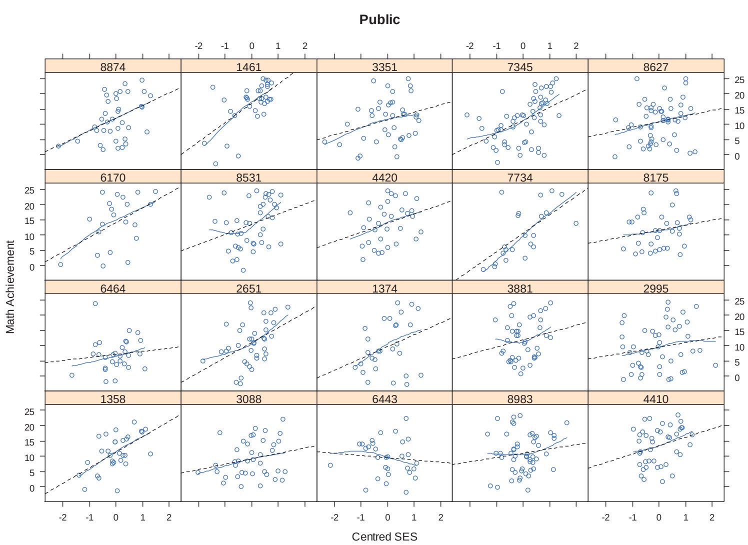

Figure 2. Math achievement by centred SES for students in 20 randomly selected public schools.

- In each scatterplot, the broken line is the linear least-squares fit to the data, while the solid line gives a nonparametric-regression fit. The number at the top of each panel is the ID number of the school.
- Particularly given the relatively small numbers of students in individual schools, the linear regressions seem to do a reasonable job of summarizing the relationship between math achievement and SES within schools.
- Although there is substantial variation in the regression lines among schools, there also seems to be a systematic difference between Catholic and public schools: The lines for the public schools appear to have steeper slopes on average.

▶ "SES" in these scatterplots is expressed as deviations from the school mean SES.
- That is, the average SES for students in a particular school is subtracted from each individual student's SES.

- *Centering* SES in this manner makes the within-school intercept from the regression of math achievement on SES equal to the average math achievement score in the school:
  – In the ith school we have the regressing equation
  $$\text{mathach}_{ij} = \alpha_{0i} + \alpha_{1i}(\text{ses}_{ij} - \overline{\text{ses}}_{i\cdot}) + \varepsilon_{ij}$$
  where
  $$\overline{\text{ses}}_{i\cdot} = \frac{\sum_{j=1}^{n_i} \text{ses}_{ij}}{n_i}$$
  – Then the least-squares estimate of the intercept is
  $$\widehat{\alpha}_{0i} = \overline{\text{mathach}}_{i\cdot} = \frac{\sum_{j=1}^{n_i} \text{mathach}_{ij}}{n_i}$$
- A more general point is that it is helpful for interpretation of hierarchical (and other!) models to scale the explanatory variables so that the parameters of the model represent quantities of interest.

- ▶ Having satisfied myself that linear regressions reasonably represent the within-school relationship between math achievement and SES, I fit this model by least squares to the data from each of the 160 schools.
  - Here are three displays of these coefficients:
    - – Figures 3 and 4 shows confidence intervals for the intercept and slope estimates for Catholic and public schools.
    - – Figure 5 shows boxplots of the intercepts and slopes for Catholic and public schools.
    - – It is apparent that the individual slopes and intercepts are not estimated very precisely, and there is also a great deal of variation from school to school.
    - – On average, however, Catholic schools have larger intercepts (i.e., a higher average level of math achievement) and lower slopes (i.e., less of a relationship between math achievement and SES).

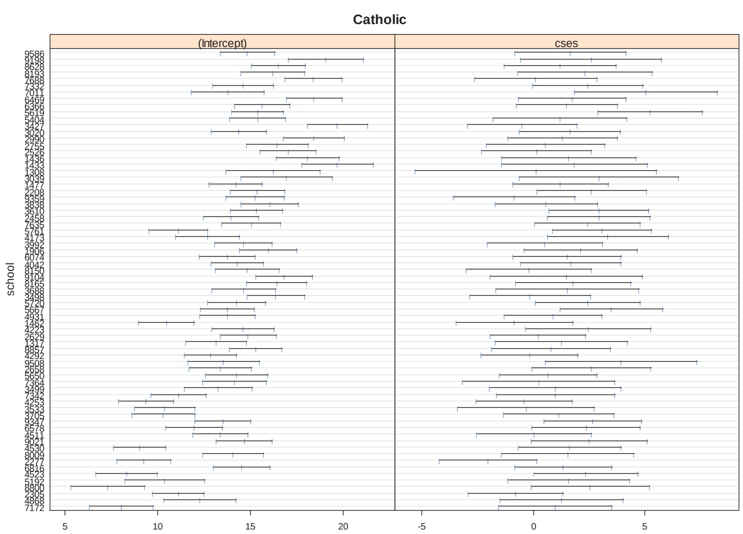

Figure 3. Confidence intervals for least-squares intercepts (left) and slopes (right) for the within-school regressions of math achievement on centered SES: 70 Catholic high schools.

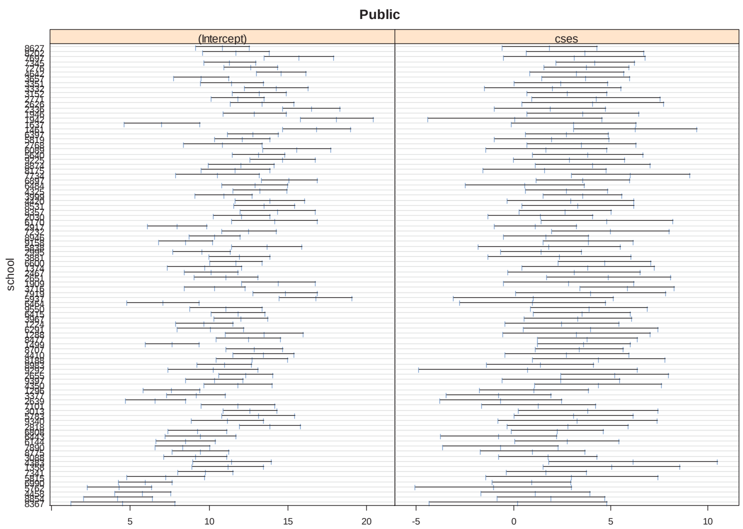

Figure 4. Confidence intervals for least-squares intercepts (left) and slopes (right) for the within-school regressions of math achievement on centered SES: 90 public high schools.

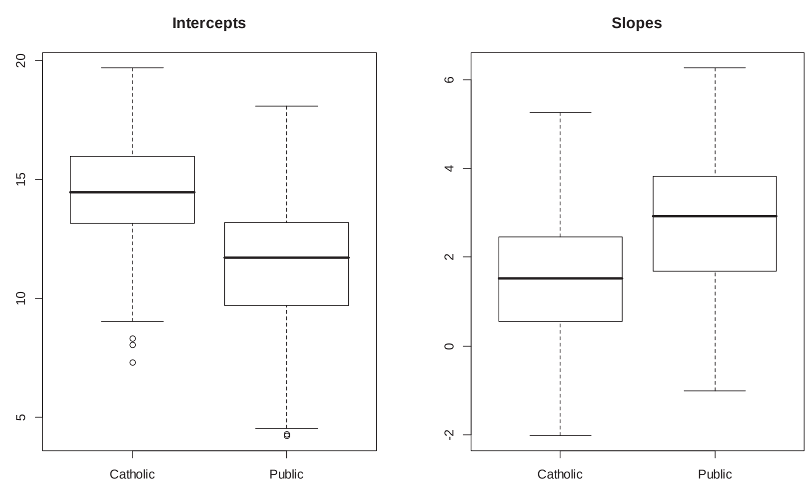

Figure 5. Boxplots of within-school coefficients for the least-squares regression of math achievement on school-centered SES, for 70 Catholic and 90 public schools.

- Figure 6 shows the relationship between the within-school intercepts and slopes and mean school SES.
  - There is a moderately strong and reasonably linear relationship between the within-school intercepts (i.e., average math achievement) and the average level of SES in the schools.
  - The slopes are weakly and, apparently, nonlinearly related to average SES.

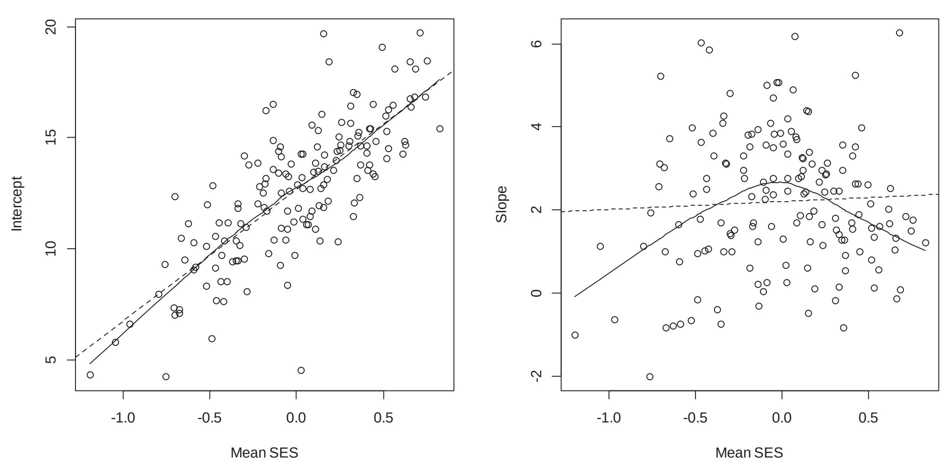

Figure 6. Within-school intercepts and slopes by mean SES. In each panel, the broken line is the linear least-squares fit and the solid line is from a nonparametric regression.

3.1 Formulating a Mixed Model

- We already have a "level-1" model for math achievement:
$$\text{mathach}_{ij} = \alpha_{0i} + \alpha_{1i}\text{cses}_{ij} + \varepsilon_{ij}$$
where $\text{cses}_{ij} = \text{ses}_{ij} - \overline{\text{ses}}_{i\cdot}$
- A "level-2" model relates the coefficients in the "level-1" model to characteristics of schools.
  - Our exploration of the data suggests the following level-2 model:
$$\begin{aligned}\alpha_{0i} &= \gamma_{00} + \gamma_{01}\overline{\text{ses}}_{i\cdot} + \gamma_{02}\text{sector}_i + u_{0i} \\ \alpha_{1i} &= \gamma_{10} + \gamma_{11}\overline{\text{ses}}_{i\cdot} + \gamma_{12}\overline{\text{ses}}^2_{i\cdot} + \gamma_{13}\text{sector}_i + u_{1i}\end{aligned}$$
where sector is a dummy variable, coded 1 (say) for Catholic schools and 0 for public schools.

- Substituting the school-level equation into the individual-level equation produces the *combined* or *composite model:*
$$\begin{aligned}\text{mathach}_{ij} &= (\gamma_{00} + \gamma_{01}\overline{\text{ses}}_{i\cdot} + \gamma_{02}\text{sector}_i + u_{0i}) \\ &\quad + \left(\gamma_{10} + \gamma_{11}\overline{\text{ses}}_{i\cdot} + \gamma_{12}\overline{\text{ses}}^2_{i\cdot} + \gamma_{13}\text{sector}_i + u_{1i}\right)\text{cses}_{ij} + \varepsilon_{ij} \\ &= \gamma_{00} + \gamma_{01}\overline{\text{ses}}_{i\cdot} + \gamma_{02}\text{sector}_i + \gamma_{10}\text{cses}_{ij} \\ &\quad + \gamma_{11}\overline{\text{ses}}_{i\cdot} \times \text{cses}_{ij} + \gamma_{12}\overline{\text{ses}}^2_{i\cdot} \times \text{cses}_{ij} + \gamma_{13}\text{sector}_i \times \text{cses}_{ij} \\ &\quad + u_{0i} + u_{1i}\text{cses}_{ij} + \varepsilon_{ij}\end{aligned}$$
  - Except for notation, this is a mixed model in Laird-Ware form, as we can see by replacing γs with βs and us with bs:
$$\begin{aligned}y_{ij} &= \beta_1 + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} \\ &\quad + \beta_5 x_{5ij} + \beta_6 x_{6ij} + \beta_7 x_{7ij} \\ &\quad + b_{1i} + b_{2i} z_{2ij} + \varepsilon_{ij}\end{aligned}$$

- Note that all explanatory variables in the Laird-Ware form of the model carry subscripts i for schools and j individuals within schools, even when the explanatory variable in question is constant within schools.
  - Thus, for example, $x_{2ij} = \overline{\text{ses}}_{i\cdot}$ (and so all individuals in the same school share a common value of school-mean SES).
  - There is both a data-management issue here and a conceptual point:
  - With respect to data management, software that fits the Laird-Ware form of the model (such as the `lme` or `lmer` functions in R) requires that level-2 explanatory variables (here sector and school-mean SES, which are characteristics of schools) appear in the level-1 (i.e., student) data set — much as the person $\times$ time-period data set that we employed in survival analysis with time-varying covariates required that time-constant covariates appear on the data record for each time period.

  - The conceptual point is that the model can incorporate *contextual effects* — characteristics of the level-2 units can influence the level-1 response variable.
  - Such contextual effects are of two kinds:
    - *Compositional effects*, such as the effect of school-mean SES, which are composed from characteristics of individuals within a level-2 unit.
    - Effects of characteristics of the level-2 units, such as school sector, that do not pertain to the level-1 units.

▶ Rather than proceeding with this relatively complicated model, let us first investigate some simpler mixed-effects models derived from it.

3.1.1 Random-Effects One-Way Analysis of Variance

- Consider the following level-1 and level-2 models:

$$\begin{aligned} \text{mathach}_{ij} &= \alpha_{0i} + \varepsilon_{ij} \\ \alpha_{0i} &= \gamma_{00} + u_{0i} \end{aligned}$$

- The combined model is

$$\text{mathach}_{ij} = \gamma_{00} + u_{0i} + \varepsilon_{ij}$$

  - In Laird-Ware form:

$$y_{ij} = \beta_1 + b_{1i} + \varepsilon_{ij}$$

- This is a *random-effects one-way ANOVA model* with one *fixed effect*, β_1, representing the general population mean of math achievement, and two *random effects*:
  - b_{1i}, representing the deviation of math achievement in school i from the general mean; that is, $\mu_i = \beta_1 + b_{1i}$ is mean math achievement in school i.

  - ε_{ij}, representing the deviation of individual j's math achievement in school i from the school mean.
  - Two observations y_{ij} and $y_{ij'}$ in school i are not independent because they share the random effect, b_{1i}.
- There are also two *variance components* for this model:
  - $\psi_1^2 = \text{Var}(b_{1i})$ is the variance among school means.
  - $\sigma^2 = \text{Var}(\varepsilon_{ij})$ is the variance among individuals in the same school.
- Because the b_{1i} and ε_{ij} are assumed to be independent, variation in math scores among individuals can be decomposed into these two variance components:

$$\text{Var}(y_{ij}) = E\left[(b_{1i} + \varepsilon_{ij})^2\right] = \psi_1^2 + \sigma^2$$

  [since $E(b_{1i}) = E(\varepsilon_{ij}) = 0$, and hence $E(y_{ij}) = \beta_1$].

- The *intra-class correlation coefficient* is the proportion of variation in individuals' scores due to differences among schools:

$$\rho = \frac{\psi_1^2}{\text{Var}(y_{ij})} = \frac{\psi_1^2}{\psi_1^2 + \sigma^2}$$

- ρ may also be interpreted as the correlation between the math scores of two individuals from the same school. That is,

$$\begin{aligned}
\text{Cov}(y_{ij}, y_{ij'}) &= E\left[(b_{1i} + \varepsilon_{ij})(b_{1i} + \varepsilon_{ij'})\right] = E(b_{1i}^2) = \psi_1^2 \\
\text{Var}(y_{ij}) &= \text{Var}(y_{ij'}) = \psi_1^2 + \sigma^2 \\
\text{Cor}(y_{ij}, y_{ij'}) &= \frac{\text{Cov}(y_{ij}, y_{ij'})}{\sqrt{\text{Var}(y_{ij}) \times \text{Var}(y_{ij'})}} = \frac{\psi_1^2}{\psi_1^2 + \sigma^2} = \rho
\end{aligned}$$

▶ The `lme` function in the **nlme** package in R provides two methods to estimate mixed-effects models (as does the `lmer` function in the **lme4** package):

- *Full maximum-likelihood (ML) estimation* of the model maximizes the likelihood with respect to all of the parameters of the model simultaneously (i.e., both the fixed-effects parameters and the variance components).
- *Restricted* (or *residual*) *maximum-likelihood (REML) estimation* integrates the fixed effects out of the likelihood and estimates the variance components; given the estimates of the variance components, estimates of the fixed effects are recovered.
- A disadvantage of ML estimates of variance components is that they are biased downwards in finite samples (much as the ML estimate of the error variance in the standard linear model is biased downwards).
- The REML estimates, in contrast, correct for loss of degrees of freedom due to estimating the fixed effects.

- The difference between the ML and REML estimates can be important when the number of "clusters" (i.e., level-2 units) in the data is small.

▶ ML and REML estimates for the current example, where there are 160 schools (level-2 units), are nearly identical:

Parameter	*ML Estimate*	*REML Estimate*
β_1	12.637	12.637
ψ_1	2.925	2.935
σ	6.257	6.257

- Note that the standard deviations (rather than the variances) of the random effects are shown.
- The estimated intra-class correlation coefficient is
$$\widehat{\rho} = \frac{2.935^2}{2.935^2 + 6.257^2} = 0.180$$
and so 18 percent of the variation in students' math-achievement scores is "attributable" to differences among schools.

3.1.2 Random-Coefficients Regression Model

▶ Let us introduce school-centered SES into the level-1 model as an explanatory variable,
$$\text{mathach}_{ij} = \alpha_{0i} + \alpha_{1i}\text{cses}_{ij} + \varepsilon_{ij}$$
and allow for random intercepts *and* slopes in the level-2 model:
$$\begin{aligned}\alpha_{0i} &= \gamma_{00} + u_{0i}\\ \alpha_{1i} &= \gamma_{10} + u_{1i}\end{aligned}$$

▶ The combined model is now
$$\begin{aligned}\text{mathach}_{ij} &= (\gamma_{00} + u_{0i}) + (\gamma_{10} + u_{1i})\,\text{cses}_{ij} + \varepsilon_{ij}\\ &= \gamma_{00} + \gamma_{10}\text{cses}_{ij} + u_{0i} + u_{1i}\text{cses}_{ij} + \varepsilon_{ij}\end{aligned}$$

- In Laird-Ware form:
$$y_{ij} = \beta_1 + \beta_2 x_{2ij} + b_{1i} + b_{2i} z_{2ij} + \varepsilon_{ij}$$

▶ This model is a *random-coefficients regression model*.

- The fixed-effects coefficients β_1 and β_2 represent, respectively, the average within-schools population intercept and slope.
  - Because SES is centered within schools, the intercept β_1 represents the general level of math achievement in the population (in the sense of the average within-school mean).
- The model has four variance-covariance components:
  - $\psi_1^2 = \text{Var}(b_{1i})$ is the variance among school intercepts (i.e., school means, because SES is school-centered).
  - $\psi_2^2 = \text{Var}(b_{2i})$ is the variance among within-school slopes.
  - $\psi_{12} = \text{Cov}(b_{1i}, b_{2i})$ is the covariance between within-school intercepts and slopes.
  - $\sigma^2 = \text{Var}(\varepsilon_{ij})$ is the error variance around the within-school regressions.

- The *composite error* for individual j in school i is
$$\zeta_{ij} = b_{1i} + b_{2i}z_{2ij} + \varepsilon_{ij}$$
  - The variance of the composite error is
$$\text{Var}(\zeta_{ij}) = E(\zeta_{ij}^2) = E\left[(b_{1i} + b_{2i}z_{2ij} + \varepsilon_{ij})^2\right] = \psi_1^2 + z_{2ij}^2\psi_2^2 + 2z_{2ij}\psi_{12} + \sigma^2$$
  - And the covariance of the composite errors for two individuals j and j' in the same school is
$$\begin{aligned}\text{Cov}(\zeta_{ij}, \zeta_{ij'}) &= E(\zeta_{ij} \times \zeta_{ij'}) = E\left[(b_{1i} + b_{2i}z_{2ij} + \varepsilon_{ij})(b_{1i} + b_{2i}z_{2ij'} + \varepsilon_{ij'})\right] \\ &= \psi_1^2 + z_{2ij}z_{2ij'}\psi_2^2 + (z_{2ij} + z_{2ij'})\psi_{12}\end{aligned}$$
  - Consequently the composite errors are heteroscedastic, and errors for individuals in the *same* school are correlated.
  - But the composite errors for two individuals in *different* schools are independent.

▶ ML and REML estimates for the model are as follows:

Parameter	ML Estimate	Std. Error	REML Estimate	Std. Error
β_1	12.636	0.244	12.636	0.245
β_2	2.193	0.128	2.193	0.128
ψ_1	2.936		2.946	
ψ_2	0.823		0.833	
ψ_{12}	0.041		0.042	
σ	6.058		6.058	

- Again, the ML and REML estimates are very close.
- Note that I've given standard errors only for the fixed effects.
  - Standard errors for variance and covariance components can be obtained in the usual manner from the inverse of the information matrix, but tests and confidence intervals based on these standard errors tend not to be accurate.

  - We can, however, test variance and covariance components by a likelihood-ratio test, contrasting the (restricted) log-likelihood for the fitted model with that for a model removing the random effects in question.
  - For example, for the current model (say model 1), removing $b_{i2}z_{2ij}$ from the model (producing, say, model 0) implies that the SES slope is identical across schools.
    - · Removing $b_{i2}z_{2ij}$ from the model gets rid of two variance-covariance parameters, ψ_2 and ψ_{12}.
    - · A likelihood-ratio test for these parameters (using REML) suggests that they should not be omitted from the model:

$$\begin{aligned}\log_e L_1 &= -23,357.12\\ \log_e L_0 &= -23,362.00\\ G^2 &= 2(\log_e L_1 - \log_e L_0) = 9.76,\ df = 2,\ p = .008\end{aligned}$$

▶ *Cautionary Remarks:*

- Because REML estimates are calculated integrating out the fixed effects, one cannot legitimately perform likelihood-ratio tests across models with *different* fixed effects when the models are estimated by REML.
  - Likelihood-ratio for variance-covariance components across nested models with identical fixed effects are perfectly fine, however.
- The null hypothesis for the likelihood-ratio test of a variance (here ψ_2^2) sets the variance to 0, which is on the boundary of the parameter space; the p-value should be adjusted for this constraint (as explained in the reading).
- A common source of estimation difficulties in mixed models is the specification of overly complex random effects.
  - Interest usually centers in the fixed effects, and it often pays to try to simplify the random-effect part of the model.

3.1.3 Coefficients-as-Outcomes Model

▶ The regression-coefficients-as-outcomes model introduces explanatory variables at level 2 to account for variation among the level-1 regression coefficients. This returns us to the model that we originally formulated for the math-achievement data:

- at level 1,

$$\text{mathach}_{ij} = \alpha_{0i} + \alpha_{1i}\text{cses}_{ij} + \varepsilon_{ij}$$

- at level 2,

$$\begin{aligned}\alpha_{0i} &= \gamma_{00} + \gamma_{01}\overline{\text{ses}}_{i\cdot} + \gamma_{02}\text{sector}_i + u_{0i}\\ \alpha_{1i} &= \gamma_{10} + \gamma_{11}\overline{\text{ses}}_{i\cdot} + \gamma_{12}\overline{\text{ses}}^2_{i\cdot} + \gamma_{13}\text{sector}_i + u_{1i}\end{aligned}$$

- The combined model:

$$\begin{aligned}\text{mathach}_{ij} &= \gamma_{00} + \gamma_{01}\overline{\text{ses}}_{i\cdot} + \gamma_{02}\text{sector}_i + \gamma_{10}\text{cses}_{ij}\\ &\quad + \gamma_{11}\overline{\text{ses}}_{i\cdot} \times \text{cses}_{ij} + \gamma_{12}\overline{\text{ses}}^2_{i\cdot} \times \text{cses}_{ij} + \gamma_{13}\text{sector}_i \times \text{cses}_{ij}\\ &\quad + u_{0i} + u_{1i}\text{cses}_{ij} + \varepsilon_{ij}\end{aligned}$$

- The combined model in Laird-Ware form:

$$\begin{aligned} y_{ij} = & \beta_1 + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} \\ & + \beta_5 x_{5ij} + \beta_6 x_{6ij} + \beta_7 x_{7ij} \\ & + b_{1i} + b_{2i} z_{2ij} + \varepsilon_{ij} \end{aligned}$$

- This model has more fixed effects than the preceding random-coefficients regression model, but the same random effects and variance components: $\psi_1^2 = \text{Var}(b_{1i})$, $\psi_2^2 = \text{Var}(b_{2i})$, $\psi_{12} = \text{Cov}(b_{1i}, b_{2i})$, and $\sigma^2 = \text{Var}(\varepsilon_{ij})$.

▶ After fitting this model to the data by REML, I tested to check whether random intercepts and slopes are still required:

Model	*Omitting*	$\log_e L$
1	—	$-23,247.70$
2	ψ_1^2, ψ_{12}	$-23,357.86$
3	ψ_2^2, ψ_{12}	$-23,247.93$

- Thus, the test for random intercepts is highly statistically significant, $G^2 = 219.86$, $df = 2$, $p \simeq 0$.
- But the test for random slopes is not, $G^2 = 0.46$, $df = 2$, $p = .80$: Apparently, the level-2 explanatory variables do a sufficiently good job of accounting for differences in slopes that the variance component for slopes is no longer needed.
- The same caveat as before applies: These p-values should be adjusted for constraining $\psi_1^2 = 0$ and $\psi_2^2 = 0$.

- Refitting the model removing $b_{i2}z_{2ij}$ produces the following REML estimates:

Parameter	*Term*	*REML Estimate*	*Std. Error*
β_1	intercept	12.128	0.199
β_2	$\overline{\text{ses}}_{i\cdot}$	5.337	0.369
β_3	sector$_i$	1.225	0.306
β_4	cses$_{ij}$	3.140	0.166
β_5	$\overline{\text{ses}}_{i\cdot} \times$ cses$_{ij}$	0.755	0.308
β_6	$\overline{\text{ses}}^2_{i\cdot} \times$ cses$_{ij}$	-1.647	0.575
β_7	sector$_i \times$ cses$_{ij}$	-1.516	0.237
ψ_1	(intercept)	1.541	
σ	(ε_{ij})	6.060	

- These estimates, all of which are statistically significant, have the following interpretations:
  - $\widehat{\beta}_1 = 12.128$ is the estimated general level of math achievement in public schools (where the dummy variable sector is coded 0) at mean school SES.
    - The interpretation of this coefficient depends upon the fact that $\overline{\text{ses}}_{i\cdot}$ (school SES) is centered to a mean of 0 across schools.
  - $\widehat{\beta}_2 = 5.337$ is the estimated increase in mean math achievement associated with a one-unit increase in school SES.
  - $\widehat{\beta}_3 = 1.225$ is the estimated difference in mean math achievement between Catholic and public schools at fixed levels of school SES.
  - $\widehat{\beta}_1$, $\widehat{\beta}_2$, and $\widehat{\beta}_3$, therefore, describe the *between-schools* regression of mean math achievement on school characteristics.

– Figure 7 shows how the coefficients $\widehat{\beta}_4$, $\widehat{\beta}_5$, $\widehat{\beta}_6$, and $\widehat{\beta}_7$ combine to produce the level-1 (i.e., *within-school*) coefficient for SES.
  · At fixed levels of school SES, individual SES is more positively related to math achievement in public than in Catholic schools.
  · The maximum positive effect of individual SES is in schools with a slightly higher than average SES level; the effect declines at low and high levels of school SES, and becomes negative at the lowest levels of school SES.

– An alternative, and more intuitive representation of the fitted model is shown in Figure 8, which graphs the fitted within-school regression of math achievement on centered SES for Catholic and public schools and for three levels of school SES: -0.7 (the approximate 5th percentile of school SES), 0 (the median), and 0.7 (the 95th percentile).

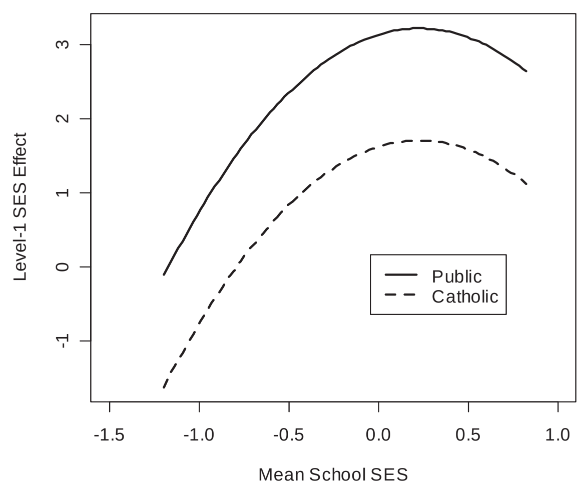

Figure 7. The level-1 effect of SES as a function of type of school (Catholic or public) and mean school SES.

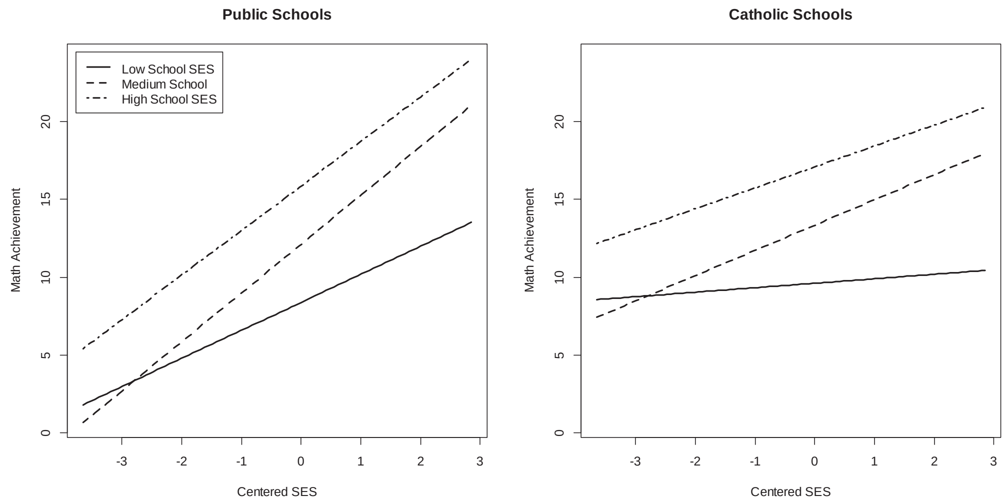

Figure 8. Fitted within-school regressions of math achievement on centred SES for public and Catholic schools at three levels of mean school SES.

4. Modeling Longitudinal Data

▶ In most respects, modeling *longitudinal data* — where there are multiple observations on individuals who are followed over time — is similar to modeling hierarchical data.

- We can think of individuals as analogous to level-2 units, and measurement occasions as analogous to level-1 units.
- Just as it is generally unreasonable to suppose in hiearchical data that observations for individuals in the same level-2 unit are independent, so is it generally unreasonable to suppose that observations taken on different occasions for the same individual are independent.
- An additional complication in longitudinal data is that it may no longer be reasonable to assume that the “level-1” errors ε_{ij} are independent, since observations taken close in time on the same individual may well be more similar than observations farther apart in time.

- When this happens, we say that the errors are *autocorrelated*.
- The linear mixed model makes provision for autocorrelated errors.

▶ In composing a mixed model for longitudinal data, we can work either with the hierarchical form of the model or with the composite (Laird-Ware) form.

▶ Consider the following example, drawn from work by Blackmore, Davis, and Fox on the exercise histories of 138 teenaged girls who were hospitalized for eating disorders and of 93 “control” subjects.

- There are several observations for each subject, but because the girls were hospitalized at different ages, the number of observations and the age at the last observation vary.

- The variables in the data set are:
  - `subject`: an identification number, necessary to keep track of which observations belong to each subject.
  - `age`: the subject’s age, in years, at the time of observation. All but the last observation for each subject were collected retrospectively at intervals of two years, starting at age eight. The age at the last observation is recorded to the nearest day.
  - `exercise`: the amount of exercise in which the subject engaged, expressed as hours per week.
  - `group`: a factor indicating whether the subject is a patient or a control.
- It is of interest here to determine the typical trajectory of exercise over time, and to establish whether this trajectory differs between eating-disordered and control subjects.
- Preliminary examination of the data suggests a log transformation of exercise.

– Because about 12 percent of the data values are 0, it is necessary to add a small constant to the data before taking logs. I used $5/60 = 1/12$ (i.e., 5 minutes).

– An alternative would be to fit a model (such as an appropriate generalized linear model) that takes explicit account of the non-negative character of the response variable.

– Figure 9, for example, shows that the original exercise scores are highly skewed, but that the log-transformed scores are much more symmetrically distributed.

- Figure 10 shows the exercise trajectories for 20 randomly selected control subjects and 20 randomly selected patients.
  – The small number of observations per subject and the substantial irregular intra-subject variation make it hard to draw conclusions, but there appears to be a more consistent pattern of increasing exercise among patients than among the controls.

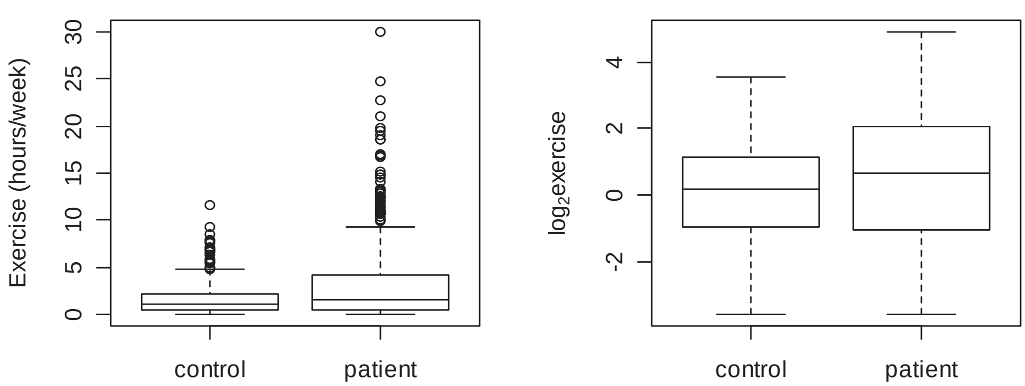

Figure 9. Boxplots of exercise and log-exercise for controls and patients, for measurements taken on all occasions. Note that logs are to the base 2.

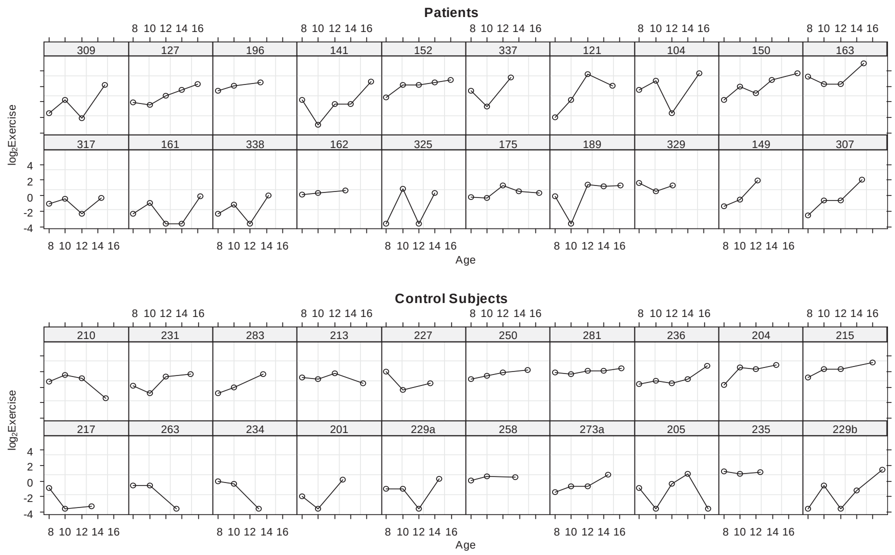

Figure 10. Exercise trajectories for 20 randomly selected patients and 20 randomly selected controls.

– With so few observations per subject, and without clear evidence that it is inappropriate, we would be loath to fit a model more complicated than a linear trend.

▶ A linear "growth curve" characterizing subject i's trajectory suggests the level-1 model

$$\log\text{-exercise}_{ij} = \alpha_{0i} + \alpha_{1i}(\text{age}_{ij} - 8) + \varepsilon_{ij}$$

- I have subtracted 8 from age, and so α_{0i} represents the level of exercise at 8 years of age — the start of the study.

▶ Our interest in detecting differences in exercise histories between subjects and controls suggests the level-2 model

$$\begin{aligned}\alpha_{0i} &= \gamma_{00} + \gamma_{01}\text{group}_i + u_{0i}\\ \alpha_{1i} &= \gamma_{10} + \gamma_{11}\text{group}_i + u_{1i}\end{aligned}$$

where group is a dummy variable coded 1 for subjects (say) and 0 for controls.

▶ Substituting the level-2 model into the level-1 model produces the combined model

$$\begin{aligned}\log\text{-exercise}_{ij} &= (\gamma_{00} + \gamma_{01}\text{group}_i + u_{0i}) \\ &\quad + (\gamma_{10} + \gamma_{11}\text{group}_i + u_{1i})(\text{age}_{ij} - 8) + \varepsilon_{ij} \\ &= \gamma_{00} + \gamma_{01}\text{group}_i + \gamma_{10}(\text{age}_{ij} - 8) \\ &\quad + \gamma_{11}\text{group}_i \times (\text{age}_{ij} - 8) + u_{0i} + u_{1i}(\text{age}_{ij} - 8) + \varepsilon_{ij}\end{aligned}$$

- or, in Laird-Ware form,

$$y_{ij} = \beta_1 + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + b_1 + b_2 z_{2ij} + \varepsilon_{ij}$$

▶ Fitting this model to the data produces the following estimates of the fixed effects and variance-covariance components:

Parameter	*Term*	*REML Estimate*	*Std. Error*
β_1	intercept	-0.2760	0.1824
β_2	group$_i$	-0.3540	0.2353
β_3	age$_{ij} - 8$	0.0640	0.0314
β_4	group$_i \times$ (age$_{ij} - 8$)	0.2399	0.0394
ψ_1	(intercept)	1.4435	
ψ_2	(age$_{ij} - 8$)	0.1648	
ψ_{12}	(intercept, age$_{ij} - 8$)	-0.0668	
σ	(ε_{ij})	1.2441	

- Letting "model 1" represent the model above, I tested whether random intercepts or random slopes could be omitted from the model:

Model	*Omitting*	$\log_e L$
1	—	-1807.07
2	ψ_1^2, ψ_{12} (random intercepts)	-1911.04
3	ψ_2^2, ψ_{12} (random slopes)	-1816.13

  – Both likelihood-ratio tests are highly statistically significant (particularly the one for random intercepts), suggesting that both random intercepts and random slopes are required (recall that the p-values should be adjusted).

▶ The model that I have fit to the Blackmore et al. data assumes independent errors, ε_{ij}.

- The *composite errors*, $\zeta_{ij} = b_1 + b_2 z_{2ij} + \varepsilon_{ij}$, are *correlated* within individuals, however, as we previously established for mixed models applied to hierarchical data.

- In the current context z_{2ij} is the time of observation (i.e., age minus eight years), and the variance and covariances of the composite residuals are (as we previously established)

$$\begin{aligned}\text{Var}(\zeta_{ij}) &= \psi_1^2 + z_{2ij}^2\psi_2^2 + 2z_{2ij}\psi_{12} + \sigma^2 \\ \text{Cov}(\zeta_{ij}, \zeta_{ij'}) &= \psi_1^2 + z_{2ij}z_{2ij'}\psi_2^2 + (z_{2ij} + z_{2ij'})\psi_{12}\end{aligned}$$

- The actual observations are not taken at entirely regular intervals, but assume that we have observations for the same individual i taken at $z_{2i1} = 0, z_{2i2} = 2, z_{2i3} = 4,$ and $z_{2i4} = 6$ (i.e., at 8, 10, 12, and 14 years of age).

  – Then the estimated covariance matrix for the composite errors is

$$\widehat{\text{Cov}}(\zeta_{i1}, \zeta_{i2}, \zeta_{i3}, \zeta_{i4}) = \begin{bmatrix} 3.631 & 1.950 & 1.816 & 1.683 \\ 1.950 & 3.473 & 1.900 & 1.875 \\ 1.816 & 1.900 & 3.532 & 2.068 \\ 1.683 & 1.875 & 2.068 & 3.808 \end{bmatrix}$$

– and the correlations for the composite errors are

$$\widehat{\text{Cor}}(\zeta_{i1}, \zeta_{i2}, \zeta_{i3}, \zeta_{i4}) = \begin{bmatrix} 1.0 & .549 & .507 & .453 \\ .549 & 1.0 & .543 & .516 \\ .507 & .543 & 1.0 & .564 \\ .453 & .516 & .564 & 1.0 \end{bmatrix}$$

– The correlations across composite errors are moderately high, and the pattern is what we would expect: The correlations tend to decline with the time-separation between occasions. This pattern, however, does not have to hold.

▶ The linear mixed model allows for correlated level-1 errors within individuals,

$$\boldsymbol{\varepsilon}_i \sim \mathbf{N}_{n_i}(\mathbf{0}, \sigma^2 \boldsymbol{\Lambda}_i)$$

- For a model with correlated errors to be identified, however, the matrix $\boldsymbol{\Lambda}_i$ cannot consist of independent parameters; instead, the elements of this matrix are expressed in terms of a much smaller number of fundamental parameters.

- For example, for equally spaced occasions, a very common model for the intra-individual errors is the *first-order autoregressive* [or *AR(1)*] *process*:

$$\varepsilon_{ij} = \phi \varepsilon_{i,j-1} + \upsilon_j$$

where

$$\upsilon_j \sim N(0, \sigma_\upsilon^2);\ \upsilon_j, \upsilon_{j'} \text{ independent for } j \neq j'$$

and

$$|\phi| < 1$$

- Then the *autocorrelation* between two errors one time-period apart (i.e., at *lag* 1) is $\rho(1) = \phi$, and the *autocorrelation* between two errors s time-periods apart (at lag s) is $\rho(s) = \phi^{|s|}$.
- Figure 11 shows two *autocorrelation functions* corresponding to first-order autoregressive processes, one for $\phi = .7$, and the other for $\phi = -.7$.
  – Note that in both cases, the autocorrelations decay as the lag grows.

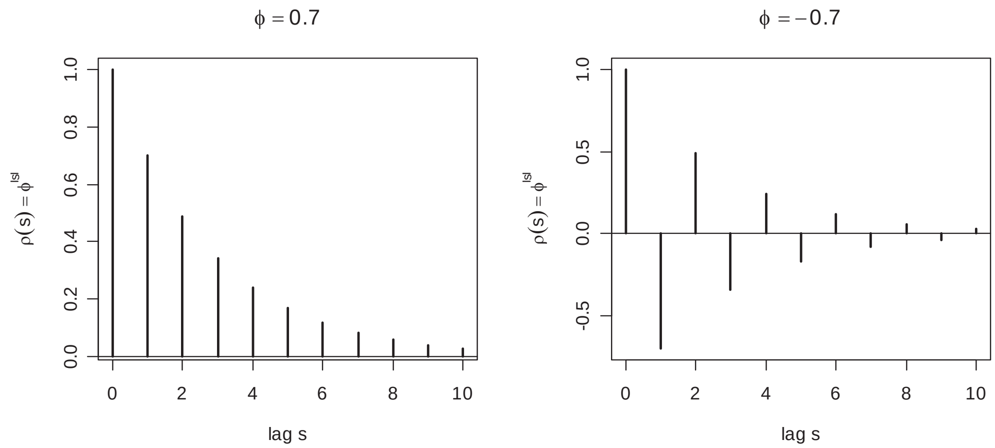

Figure 11. Autocorrelation functions for $\phi = .7$ (left) and $\phi = -.7$ (right).

- The `lme` function in R provides several other time-series error processes for equally spaced observations besides AR(1), as well as the possibility of adding still more such processes.
- The occasions for the Blackmore et al. data are not equally spaced, however.
  - For data such as these, `lme` provides a *continuous first-order autoregressive process*, with the property that
  $$\text{corr}(\varepsilon_{it}, \varepsilon_{i,t+s}) = \rho(s) = \phi^{|s|}$$
  where the time-interval between observations, s, need not be an integer.

▶ I tried to fit the same mixed-effects model to the data as before, except allowing for first-order autoregressive level-1 errors.

- The estimation process did not converge; a closer inspectation suggests that the model has redundant parameters.
- I then fit two additional models, retaining autocorrelated within-subject errors, but omitting in turn random slopes and random intercepts.

- These models are not nested, so they cannot be compared via likelihood-ratio tests, but we can still compare the fit of the models to the data:

Model	*Log-Likelihood*	df
Independent within-subject errors, random intercepts and slopes	-1807.068	8
Correlated within-subject errors, random intercepts	-1795.484	7
Correlated within-subject errors, random slopes	-1802.294	7

  – Thus, the random-intercept model with autocorrelated within-subject errors produces the best fit to the data.
  – Trading-off parameters for the dependence of the within-subject errors against random effects is a common pattern: All three models produce similar estimates of the fixed effects.

- Estimates for a final model, incorporating random intercepts and autocorrelated errors, are as follows:

Parameter	*Term*	*REML Estimate*	*Std. Error*
β_1	intercept	-0.3070	0.1895
β_2	group$_i$	-0.2838	0.2447
β_3	age$_{ij} - 8$	0.0728	0.0317
β_4	group$_i \times ($age$_{ij} - 8)$	0.2274	0.0397
ψ_1	(intercept)	1.1497	
σ	(ε_{ij})	1.5288	
ϕ	(error autocorrelation at lag 1)	0.6312	

- Notice that the slope for the control group ($\widehat{\beta}_3$) is statistically significant, and the differences in slopes between the patient group and the controls ($\widehat{\beta}_4$) is highly statistically significant.
- The initial difference between the groups (i.e., $\widehat{\beta}_2$, the estimated difference at age 8) is non-significant.

- A graph showing the fit of the model, translating back from log-exercise to the exercise scale, appears in Figure 12.

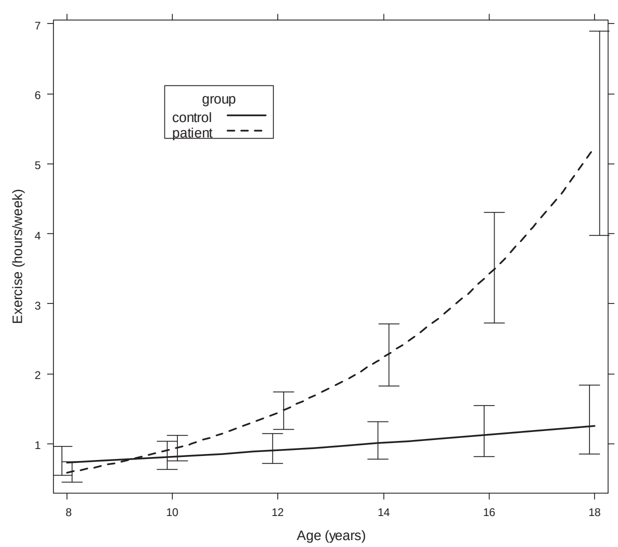

Figure 12. Fitted exercise as a function of age and group: Average trajectories based on fixed effects. The error bars show 95% confidence intervals around the fit.

5. Generalized Linear Mixed-Effects Models

5.1 Quick Review of Generalized Linear Models

▶ A *generalized linear model* consists of three components:

1. A *random component*, specifying the conditional distribution of the response variable, y_i, given the predictors. Traditionally, the random component is an *exponential family* — the normal (Gaussian), binomial, Poisson, gamma, or inverse-Gaussian.
2. A linear function of the regressors, called the *linear predictor*,
$$\eta_i = \beta_1 + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$$
on which the expected value μ_i of y_i depends.
3. A *link function* $g(\mu_i) = \eta_i$, which transforms the expectation of the response to the linear predictor. The inverse of the link function is called the *mean function*: $g^{-1}(\eta_i) = \mu_i$.

▶ In the following table, the logit, probit and complementary log-log links are for binomial or binary data:

Link	$\eta_i = g(\mu_i)$	$\mu_i = g^{-1}(\eta_i)$
identity	μ_i	η_i
log	$\log_e \mu_i$	e^{η_i}
inverse	μ_i^{-1}	η_i^{-1}
inverse-square	μ_i^{-2}	$\eta_i^{-1/2}$
square-root	$\sqrt{\mu_i}$	η_i^2
logit	$\log_e \dfrac{\mu_i}{1-\mu_i}$	$\dfrac{1}{1+e^{-\eta_i}}$
probit	$\Phi(\mu_i)$	$\Phi^{-1}(\eta_i)$
complementary log-log	$\log_e[-\log_e(1-\mu_i)]$	$1-\exp[-\exp(\eta_i)]$

- In R, generalized linear models are fit with the `glm` function, and most of the arguments of `glm` are similar to those of `lm`:
  - The response variable and regressors are given in a model `formula`
  - `data`, `subset`, and `na.action` arguments determine the data on which the model is fit.
  - The additional `family` argument is used to specify a family-generator function, which may take other arguments, such as a link function.

- The following table gives family generators and default ("canonical") links:

Family	*Default Link*	*Range of* y_i	Var$(y_i \vert \eta_i)$
`gaussian`	`identity`	$(-\infty, +\infty)$	ϕ
`binomial`	`logit`	$\frac{0, 1, ..., n_i}{n_i}$	$\mu_i(1-\mu_i)$
`poisson`	`log`	$0, 1, 2, ...$	μ_i
`Gamma`	`inverse`	$(0, \infty)$	$\phi\mu_i^2$
`inverse.gaussian`	`1/mu^2`	$(0, \infty)$	$\phi\mu_i^3$

  - For distributions in the exponential families, the conditional variance of y_i is a function of the mean μ_i and a dispersion parameter ϕ (fixed to 1 for the binomial and Poisson distributions).

- The following table shows the links available for each family in R, with the default links in bold:

	link			
family	identity	inverse	sqrt	1/mu^2
gaussian	**X**	X		
binomial				
poisson	X		X	
Gamma	X	**X**		
inverse.gaussian	X	X		**X**
quasi	**X**	X	X	X
quasibinomial				
quasipoisson	X		X	

	link			
family	log	logit	probit	cloglog
gaussian	X			
binomial	X	**X**	X	X
poisson	**X**			
Gamma	X			
inverse.gaussian	X			
quasi	X	X	X	X
quasibinomial		**X**	X	X
quasipoisson	**X**			

- The `quasi`, `quasibinomial`, and `quasipoisson` family generators do not correspond to exponential families.
  - The quasi-binomial and quasi-Poisson families can be used to fit "over-dispersed" binomial and Poisson GLMs.
  - Such models are estimated by *quasi-likelihood* methods (specifying the variance as a function of the mean and a dispersion parameter).

5.2 The Generalized Linear Mixed Model

- The *generalized linear mixed-effects model* (*GLMM*) is a straightforward extension of the generalized linear model, adding random effects to the linear predictor, and expressing the expected value of the response conditional on the random effects:

$$\begin{aligned} g(\mu_{ij}) &= g[E(y_{ij}|b_{1i}, \ldots, b_{qi})] = \eta_{ij} \\ \eta_{ij} &= \beta_1 + \beta_2 x_{2ij} + \cdots + \beta_p x_{pij} + b_{1i} z_{1ij} + \cdots + b_{qi} z_{qij} \end{aligned}$$

  - The link function $g()$ is as in generalized linear models.
  - The conditional distribution of y_{ij} given the random effects is a member of an exponential family, or — for quasi-likelihood estimation — the variance of $y_{ij}|b_{1i}, \ldots, b_{qi}$ is a function of μ_{ij} and a dispersion parameter ϕ.
  - We make the usual assumptions about the random effects: That they are multinormally distributed with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Psi}$.

- The estimation of generalized linear mixed models by ML is not so straightforward, because the likelihood function includes integrals that are analytically intractable.
  - There are several practical approaches to estimating GLMMs that involve approximating the likelihood.
  - The `glmer` function in the `lme4` package provides a generally accurate approximate ML solution.

5.3 Example: Migraine Headaches

- This example is borrowed from a paper by Kostecki-Dillon, Monette, and Wong (1999).
- In an effort to reduce the severity and frequency of migraine headaches through the use of biofeedback training, longitudinal data were collected on 133 migraine-headache sufferers.
  - The patients were each given four weekly sessions of biofeedback training.
  - They were asked to keep daily logs of their headaches for a period of 30 days prior to training, during training, and post-training, to 100 days after training began.
  - Compliance was low: e.g., only 55 patients kept a log prior to training.
  - On average, subjects recorded information on 31 days, with the number of days ranging from 7 to 121.

- Subjects were divided into three *self-selected* groups:
  (a) those who discontinued their migraine medication during the training and post-training phase of the study;
  (b) those who continued their medication, but at a reduced dose; and
  (c) those who continued their medication at the previous dose.
- I will use a binary logit mixed-effects model to analyze the incidence of headaches during the period of the study.
  - Examination of the data suggested that the incidence of headaches was invariant during the pre-training phase of the study, increased at the start of training, and then declined at a decreasing rate.
  - I decided to fit a linear trend prior to the start of training (before time 0), possibly to capture a trend that I failed to detect in exploring the data, and to transform time at day 1 and later ("time post-treatment") by taking the square-root.

- The model includes
  - an intercept, representing the level of headache incidence at the end of the pre-training period;
  - a dummy regressor coded 1 post-treatment, and 0 pre-treatment, to capture the anticipated increase in headache incidence at the start of training;
  - two dummy regressors for levels of medication; and
  - interactions between medication and treatment, and between medication and the pre- and post-treatment time trends.

- Thus, the fixed-effects part of the model is

$$\begin{aligned}\text{logit}(\pi_{ij}) = {} & \beta_1 + \beta_2 m_{1i} + \beta_3 m_{2i} + \beta_4 p_{ij} + \beta_5 t_{0ij} + \beta_6 \sqrt{t_{1ij}} \\ & + \beta_7 m_{1i} p_{ij} + \beta_8 m_{2i} p_{ij} + \beta_9 m_{1i} t_{0ij} + \beta_{10} m_{2i} t_{0ij} \\ & + \beta_{11} m_{1i} \sqrt{t_{1ij}} + \beta_{12} m_{2i} \sqrt{t_{1ij}}\end{aligned}$$

  - π_{ij} is the probability of a headache for individual $i = 1, \ldots, 133$, on occasion $j = 1, \ldots, n_i$;
  - m_{1i} is a dummy regressor coded 1 if individual i continued taking migraine medication at a reduced dose post-treatment, and m_{2i} is a dummy regressor coded 1 if individual i continued taking medication at its previous dose post-treatment;
  - p_{ij} is a dummy regressor coded 1 post-treatment (i.e., after time 0) and 0 pre-treatment;

  - t_{0ij} is time (in days) pre-treatment, running from -29 through 0, and coded 0 after treatment began; and
  - t_{1ij} is time (in days) post-treatment, running from 1 through 99, and coded 0 pre-treatment.

- I experienced a bit of difficulty fitting this model: Even without temporal autocorrelation, the random effects are complex for a small data data set.
- Type-II Wald tests for terms in the model (obtained from the `Anova` function in the **car** package)

Term	*Wald* χ^2	df	p
Medication (m_1, m_2)	22.34	2	$< .0001$
Treatment (p)	13.32	1	$< .001$
Pre-treatment trend (t_0)	0.38	1	$.54$
Post-treatment trend ($\sqrt{t_1}$)	34.60	1	$\ll .0001$
Medication $\times$ Treatment	2.38	2	$.30$
Medication $\times$ Pre-treatment	1.86	2	$.39$
Medication $\times$ Post-treatment	0.06	2	$.97$

- Tests of random effects:
  - In each case, one variance and three covariance components are removed.
  - Recall that the p-values should be adjusted for testing that a variance component is 0.

Random Effect Removed	G^2	*p*
Intercept	19.70	$.0006$
Treatment	12.08	$.017$
Pre-treatment trend	5.79	$.21$
Post-treatment trend	16.21	$.0027$

- Based on the tests for the fixed and random effects, I fit a final model that eliminates the fixed-effect interactions with medication and the pre-treatment trend fixed and random effects, obtaining these estimates:

Term	*Parameter*	*Estimate*	*Std. Error*
Intercept	β_1	-0.246	0.344
	ψ_1	1.304	–
Medication			
reduced	β_2	2.050	0.468
continuing	β_3	1.155	0.384
Treatment	β_4	1.061	0.244
	ψ_2	1.309	–
Post-treatment trend	β_6	-0.268	0.045
	ψ_4	0.239	–

- Figure 13 shows the estimated fixed effects plotted on the probability scale.

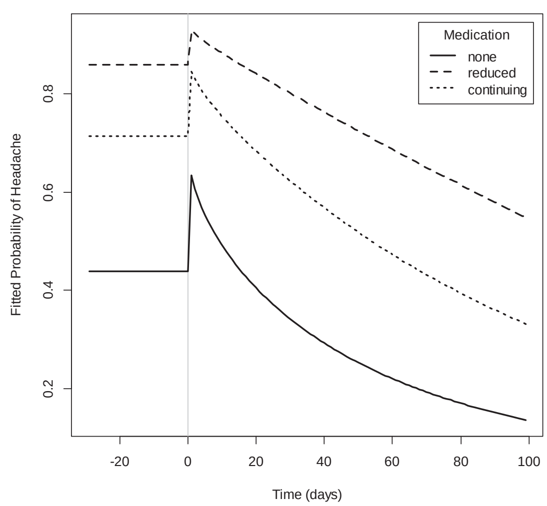

Figure 13. Fixed effects from the GLMM fit to the migraine-headaches data, with the fitted response shown on the probability scale.

