Hierarchical Models - Statistical Methods

Sarah Filippi\(^1\)
University of Oxford

Hilary Term 2016

\(^1\)With grateful acknowledgements to Neil Laws for his notes from 2012–13.
Information

Webpage: http://www.stats.ox.ac.uk/~filippi/

Lectures:
  • Week 1 – Thursday 12, Friday 11
  • Week 2 – Thursday 12, Friday 11
  • Week 3 – Thursday 12

Practical:
  • Week 3 – Friday
Outline

Hierarchical models: frequentist and Bayesian approaches.

1. Introduction of the concepts through examples
2. Inferring the parameters of a hierarchical linear model: the frequentist approach
3. Inferring the parameters of a hierarchical linear model: the Bayesian approach
4. Studying inferred Hierarchical Linear Models and cross-level interactions
5. Summary and Further properties of hierarchical models

We will rely on material that you have covered in: linear models, Bayesian statistics, R programming (e.g. lattice graphics), ...
References

1. Introduction of Hierarchical Models through examples
Rail: simple example of random effects

Example from Pinheiro and Bates (2000).

- Six railway rails were chosen at random and tested three times each.
- The Rail dataset records the time it took for an ultrasonic wave to travel the length of each rail.
- This is a one-way classification: each observation is classified according to the rail on which it was made.
## if necessary first install the package MEMSS, 
## either using the Packages menu, 
## or alternatively using something like: 
## install.packages("MEMSS")

```r
> data(Rail, package = "MEMSS")
> Rail
   Rail travel
   1 A  55
   2 A  53
   3 A  54
   4 B  26
   ....
   17 F  85
   18 F  83
```

Think of the measurements made on the same rail as a “group” – so we have 6 groups (labelled as A,...,F) with 3 measurements in each group.
The original experimenters were interested in

- the travel time for a typical rail (expected travel time)
- the variation in times between different rails (between-group variability)
- the variation in times for a single rail (within-group variability).
library(lattice)
myrail <- with(Rail, reorder(Rail, travel))
dotplot(myrail ~ travel, data = Rail,
       xlab = "Travel time (ms)", ylab = "Rail")

Between-group variability is much larger than within-group variability.
One possible model is

\[ y_{ij} = \beta_j + \epsilon_{ij}, \quad i = 1, 2, 3, \quad j = 1, \ldots, 6 \]

where

- \( i \) indexes the observations within a group
- \( j \) indexes the different groups
- \( y_{ij} \) is the travel time for observation \( i \) in group \( j \)
- the parameters \( \beta_1, \ldots, \beta_6 \) are fixed unknown constants: \( \beta_j \) is the effect of group \( j \)
- the \( \epsilon_{ij} \) are independent \( \mathcal{N}(0, \sigma^2) \) random errors.
However, this model only models the 6 specific rails that were actually measured in this experiment,

- it does not model the population of rails from which the sample was drawn
- it does not provide an estimate of the between-rail variability
- the number of parameters of the model increases linearly with the number of rails (for $N$ rails we would have $N$ parameters).
A hierarchical model treats the rail effects as random variations about an underlying population mean, as follows. We suppose that

\[ y_{ij} = \beta + b_j + \epsilon_{ij} \]

where

- \( \beta \) is a fixed unknown parameter representing the expected travel time across the whole population of rails
- \( b_j \) is a **random variable** representing the deviation from the population mean for rail \( j \)
- \( \epsilon_{ij} \) is a random variable representing the additional deviation for the \( i \)th observation on rail \( j \).
• We assume that, independently,

\[ b_j \sim N(0, \sigma_b^2) \quad \text{and} \quad \epsilon_{ij} \sim N(0, \sigma^2). \]

So \( \sigma_b^2 \) denotes the between-group variability, and \( \sigma^2 \) denotes the within-group variability.

• The parameter \( \beta \) is called a \textit{fixed} effect, and the \( b_j \) are called \textit{random} effects. A model with both types of effect is often called a \textit{mixed-effects} model.

• The model above is a mixed-model: it has a simple fixed part (just an intercept), later examples will have more interesting fixed parts; it also has a random part (the \( b_j \)).
The model has two sources of random variation:
- one of these sources is at the individual observation level \( (\epsilon_{ij}) \)
- the other source is at the group level \( (b_j) \).

Remark: No matter how many rails we have, we will always have just 3 parameters: \( \beta, \sigma_b^2 \) and \( \sigma^2 \).
Vocabulary

- Since we have two sources of variation, at different levels, such models are called *hierarchical* models and also *multilevel* models.
- Models with more than two levels of variation are also possible.
- Such models are also called *mixed-effects model*.
- Another name is *variance components*.
In our model $y_{ij} = \beta + b_j + \epsilon_{ij}$

- observations made on different groups are independent
- but observations within the same group share the same $b_j$ and are correlated.

The variance of each $y_{ij}$ is

$$\text{var}(y_{ij}) = \text{var}(b_j) + \text{var}(\epsilon_{ij}) = \sigma_b^2 + \sigma^2$$

and the covariance between observations $i$ and $i'$ (where $i \neq i'$) in group $j$ is

$$\text{cov}(y_{ij}, y_{i'j}) = \text{var}(b_j) = \sigma_b^2.$$ 

So the correlation between two observations from the same group is

$$\frac{\sigma_b^2}{\sigma_b^2 + \sigma^2}.$$
Summary

The rail example allowed us to

- introduce the concept of between-group and within-group variability
- describe how to include random-effects in simples constant models
- define hierarchical, multi-level, mixed-effect ... models.

Multilevel structures can be added to any type of regression model including linear, logistic and generalised linear models.

In this lecture, we will mainly focus on Hierarchical linear models.
Orthodont: simple linear growth curves

Example from Pinheiro and Bates (2000).

- The Orthodont dataset is a set of measurements of the distance from the pituitary gland to the pterygomaxillary fissure taken every two years from age 8 until age 14 on a sample of 27 children – 16 males and 11 females.
- Measurements of this type are sometimes referred to as growth curve data, or repeated measures or longitudinal or panel data.
- For now, we restrict to the data from the female subjects.
- Think of the measurements made on the same subject as a group.
```r
> data(Orthodont, package = "MEMSS")
> Orthodont$mySubj <- with(Orthodont,
                    reorder(Subject, distance))
> orthf <- Orthodont[Orthodont$Sex == "Female", ]

> orthf

<table>
<thead>
<tr>
<th></th>
<th>distance</th>
<th>age</th>
<th>Subject</th>
<th>Sex</th>
<th>mySubj</th>
</tr>
</thead>
<tbody>
<tr>
<td>65</td>
<td>21.0</td>
<td>8</td>
<td>F01</td>
<td>Female</td>
<td>F01</td>
</tr>
<tr>
<td>66</td>
<td>20.0</td>
<td>10</td>
<td>F01</td>
<td>Female</td>
<td>F01</td>
</tr>
<tr>
<td>67</td>
<td>21.5</td>
<td>12</td>
<td>F01</td>
<td>Female</td>
<td>F01</td>
</tr>
<tr>
<td>68</td>
<td>23.0</td>
<td>14</td>
<td>F01</td>
<td>Female</td>
<td>F01</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>108</td>
<td>28.0</td>
<td>14</td>
<td>F11</td>
<td>Female</td>
<td>F11</td>
</tr>
</tbody>
</table>
```
xyplot(distance ~ age | mySubj, data = orthf,
  aspect = "xy", type = c("g", "p", "l"),
  scales = list(x = list(at = c(8, 10, 12, 14))))
We consider the model

\[ y_{ij} = \beta_1 + b_j + \beta_2 x_{ij} + \epsilon_{ij} \]

where \( b_j \sim N(0, \sigma^2_b) \) and \( \epsilon_{ij} \sim N(0, \sigma^2) \).

Here

- \( y_{ij} \) is the distance for observation \( i \) on subject \( j \)
- the overall intercept for subject \( j \) is made up of a fixed part, the parameter \( \beta_1 \), plus a random deviation \( b_j \) representing the deviation from this fixed intercept for subject \( j \)
- \( x_{ij} \) is the age of the subject when the \( i \)th observation is made on subject \( j \).

So we are regressing distance on age, and allowing a random intercept for each subject. As before, observations on different groups (i.e. subjects) are independent, but observations within the same group are dependent.
• This model assumes that the effect of age (the growth rate) is the same for all subjects.

• We can fit a model with random effects for both the intercept and the effect of age:

\[ y_{ij} = \beta_1 + b_j + c_j x_{ij} + \epsilon_{ij} \]

where \( b_j \sim N(0, \sigma_b^2) \), \( c_j \sim N(0, \sigma_c^2) \) and \( \epsilon_{ij} \sim N(0, \sigma^2) \).

• The random part is an intercept term plus an age term, conditional on subject. That is, we have a random growth rate, in addition to a random intercept, for each subject.
Within-subject: 11 simple linear regressions, one per subject (more sensitivity to extreme observations than the mixed model)
Population: one fixed model for the whole population
Example from Gellman and Hill (2007)

• Consider an educational study with data from students in many schools \( j = 1 \ldots J \)

• We want to predict the students’ grades \( y_{ij} \) on a standardized test given their scores on a pre-test \( x_{ij} \).

• A separate regression model can be fit within each school, and the parameters from these schools can themselves be modeled as depending on school characteristics (such as the socioeconomic status of the school’s neighborhood, whether the school is public or private, and so on).

• We can otherwise build a multilevel regression model

• The student-level regression and the school-level regression here are the two levels of a multilevel model.
Models for regression coefficients

- **Varying-intercept model:** the regressions have the same slope in each of the schools, and only the intercepts vary.

\[ y_{ij} = \beta_1 + b_j + \beta_2 x_{ij} + \epsilon_{ij} \]

where \( b_j \sim N(0, \sigma_b^2) \) and \( \epsilon_{ij} \sim N(0, \sigma^2) \)

- **Varying-slope model:** the regressions have the same intercepts in each of the schools, and only the slopes vary.

\[ y_{ij} = \beta_1 + c_j x_{ij} + \epsilon_{ij} \]

where \( c_j \sim N(0, \sigma_c^2) \) and \( \epsilon_{ij} \sim N(0, \sigma^2) \)

- **Varying-intercept, varying-slope model:** intercepts and slopes both can vary by school.

\[ y_{ij} = \beta_1 + b_j + c_j x_{ij} + \epsilon_{ij} \]

where \( b_j \sim N(0, \sigma_b^2) \), \( c_j \sim N(0, \sigma_c^2) \) and \( \epsilon_{ij} \sim N(0, \sigma^2) \)
Varying intercepts

Varying slopes

Varying intercepts and slopes
Some other examples

From Gellman and Hill’s research (2007)

- Combining information for local decisions: home radon measurement and remediation
- Modeling correlations: forecasting presidential elections
- Small-area estimation: state-level opinions from national polls
- Social science modeling: police stops by ethnic group with variation across precincts
Examples:

<table>
<thead>
<tr>
<th>macro-level</th>
<th>micro-level</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i.e. groups)</td>
<td>(i.e. observations)</td>
</tr>
<tr>
<td>schools</td>
<td>teachers</td>
</tr>
<tr>
<td>classes</td>
<td>pupils</td>
</tr>
<tr>
<td>neighborhoods</td>
<td>families</td>
</tr>
<tr>
<td>districts</td>
<td>voters</td>
</tr>
<tr>
<td>firms</td>
<td>employees</td>
</tr>
<tr>
<td>families</td>
<td>children</td>
</tr>
<tr>
<td>doctors</td>
<td>patients</td>
</tr>
<tr>
<td>subjects</td>
<td>measurements</td>
</tr>
</tbody>
</table>
• So we are thinking about data with a nested structure: e.g.
  • students within schools
  • voters within districts.
• More than two levels are possible: e.g. students within classes
  within schools.
• Multilevel analysis is a suitable approach to take into account
  the social contexts as well as the individual respondents or
  subjects.
• The hierarchical linear model is a type of regression model for
  multilevel data where the dependent variable is at the lowest
  level.
• Explanatory variables can be defined at any level (including
  aggregates of micro-level variables, e.g. the average income of
  voters in a district).
2. Inferring the parameters of a hierarchical linear model: the frequentist approach
Three types of hierarchical linear models

- **Varying-intercept model**: the regressions have the same slope in each of the groups, and only the intercepts vary.

  \[ y_{ij} = \beta_1 + b_j + \beta_2 x_{ij} + \epsilon_{ij} \]

- **Varying-slope model**: the regressions have the same intercepts in each of the groups, and only the slopes vary.

  \[ y_{ij} = \beta_1 + (\beta_2 + c_j) x_{ij} + \epsilon_{ij} \]

- **Varying-intercept, varying-slope model**: intercepts and slopes both can vary by group.

  \[ y_{ij} = \beta_1 + b_j + (\beta_2 + c_j) x_{ij} + \epsilon_{ij} \]

where \( b_j \sim N(0, \sigma_b^2) \), \( c_j \sim N(0, \sigma_c^2) \) and \( \epsilon_{ij} \sim N(0, \sigma^2) \)
Matrix representation of hierarchical linear models

[Remember: we can write a linear regression model in matrix form as \( y = X\beta + \epsilon \).]

Denote by:

- \( y_j \) the vector of observations in group \( j \)
- \( \beta \) the vector of fixed parameters (with components \( \beta_1, \beta_2, \ldots \))
- \( b_j \) the vector of random effects for group \( j \) (with components \( b_{1j}, b_{2j}, \ldots \))
- \( \epsilon_j \) the vector of random errors for group \( j \).

Then, any hierarchical linear model (with one level of grouping) can be written as

\[
y_j = X_j\beta + Z_j b_j + \epsilon_j
\]

where the matrices \( X_j \) and \( Z_j \) are the known fixed-effects and random-effects regressor matrices.
The following varying-intercept model

\[ y_{ij} = b_{1j} + \beta_1 x_{ij} + \epsilon_{ij} \]

(for \( j = 1, \ldots, J \) and \( i = 1, \ldots, n_j \)) can be written in matrix form:

\[
y_j = \begin{pmatrix} y_{1j} \\ y_{2j} \\ \vdots \\ y_{njj} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} b_{1j} + \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{njj} \end{pmatrix} \beta_1 + \begin{pmatrix} \epsilon_{1j} \\ \epsilon_{2j} \\ \vdots \\ \epsilon_{njj} \end{pmatrix} = X_j \beta + Z_j b_j + \epsilon_j
\]

where \( b_j = b_{1j} \), \( \beta = \beta_j \), \( X_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{njj} \end{pmatrix} \) and \( Z_j = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \).
Similarly, the varying-intercept varying-slope model

\[ y_{ij} = \beta_1 + \beta_2 x_{ij} + b_{1j} + b_{2j} x_{ij} + \epsilon_{ij} \]

(for \( j = 1, \ldots, J \) and \( i = 1, \ldots, n_j \)) can be written in matrix form:

\[
\begin{align*}
    y_j &= \begin{pmatrix} y_{1j} \\ y_{2j} \\ \vdots \\ y_{n_jj} \end{pmatrix} = \begin{pmatrix} 1 & x_{1j} \\ 1 & x_{2j} \\ \vdots & \vdots \\ 1 & x_{n_jj} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} 1 & x_{1j} \\ 1 & x_{2j} \\ \vdots & \vdots \\ 1 & x_{n_jj} \end{pmatrix} \begin{pmatrix} b_{1j} \\ b_{2j} \end{pmatrix} + \begin{pmatrix} \epsilon_{1j} \\ \epsilon_{2j} \\ \vdots \\ \epsilon_{n_jj} \end{pmatrix} \\
    &= X_j \beta + Z_j b_j + \epsilon_j
\end{align*}
\]

Here:

\[
X_j = Z_j = \begin{pmatrix} 1 & x_{1j} \\ 1 & x_{2j} \\ \vdots & \vdots \\ 1 & x_{n_jj} \end{pmatrix}
\]
Some dimensions

- Let the groups be \( j = 1, \ldots, J \), and assume we have \( n_j \) observations in group \( j \). We can take

\[
y_j = X_j \beta + Z_j b_j + \epsilon_j \quad j = 1, \ldots, J
\]

as a general form of the \textit{hierarchical linear model}, provided that we allow \( \beta \) and \( b_j \) to have any dimensions.

- Let \( \beta \) contain \( p \) parameters, and let each \( b_j \) be a random vector of length \( q \).

- The matrices \( X_j \) and \( Z_j \) are known fixed-effects and random-effects regressor matrices:
  - the columns of \( X_j \) are the values of the explanatory variables for group \( j \) (\( X_j \) is \( n_j \times p \))
  - The columns of \( Z_j \) are usually a subset of the columns of \( X_j \) (\( Z_j \) is \( n_j \times q \))
Typically:

- We assume that the components of the vector $\epsilon_j$ are i.i.d.

$$\epsilon_j \sim \mathcal{N}(0, \sigma^2 I) .$$

- The vectors $b_j$ are assumed to be independent and normally distributed with mean 0:

$$b_j \sim \mathcal{N}(0, \Psi)$$

- However we do not assume that the components of $b_j$ are independent – we make no assumptions about the correlations of the components of $b_j$, (these are estimated).

- $b_j$ and $\epsilon_j$ are mutually independent and independent across groups

We aim at estimating the parameters $\beta$, $\sigma^2$ and $\Psi$.

Here we focus on a model with a ”single level of grouping”. See Pinheiro and Bates (2000) for details and more general case.
Likelihood function

The *likelihood function* is the probability density of the data given the parameters, regarded as a function of the parameters:

\[
L(\beta, \sigma^2, \Psi; y) = p(y|\beta, \sigma^2, \Psi).
\]

Because the random-effect variables \(b_j, j = 1 \ldots J\), are part of the model, we must integrate them out:

\[
L(\beta, \sigma^2, \Psi; y) = \prod_{j=1}^{J} p(y_j|\beta, \sigma^2, \Psi) \quad = \prod_{j=1}^{J} \int p(y_j|b_j, \beta, \sigma^2, \Psi)p(b_j|\Psi)db_j
\]

Which assumption have we used here?
Given that

\[ y_j | b_j, \beta, \sigma^2 \sim \mathcal{N}(X_j \beta + Z_j b_j, \sigma^2 I) \]

and

\[ b_j \sim \mathcal{N}(0, \Psi) \]

we obtain

\[
L(\beta, \sigma^2, \Psi; y) = \prod_{j=1}^{J} \int \frac{1}{(2\pi \sigma^2)^{n_j/2}} \exp \left( -\frac{1}{2\sigma^2} \| y_j - X_j \beta - Z_j b_j \|^2 \right) \left( \frac{1}{(2\pi)^{q/2}} \frac{1}{\sqrt{\det \Psi}} \exp \left( -b_j^T \Psi^{-1} b_j \right) \right) db_j
\]
Maximum Likelihood (ML)

Maximise the likelihood function by:

1. maximising it with respect to $\beta$ and $\sigma^2$ (similar to linear regression model),

2. replacing $\beta$ and $\sigma^2$ by their estimates in the likelihood function,

3. maximising the remaining objective function (called profile likelihood) with respect to $\Psi$.

Remarks:

- Maximum likelihood estimates of variance components are biased and tend to underestimate these parameters
- If you are interested in estimating components of variance then it is better to use another methods (see next slide).
Restricted Maximum Likelihood (REML)

- Also called "Reduced" or "Residual" Maximum Likelihood
- Maximise the following quantity

\[ L_R(\Psi, \sigma^2; y) = \int L(\beta, \Psi, \sigma^2; y) d\beta \]

1. by treating \( \sigma^2 \) as if known and maximizing objective function with respect to \( \Psi \):

\[ \widehat{\Psi}_\sigma = \arg\max_{\Psi} L_R(\Psi, \sigma^2; y) \]

2. using the estimate of \( \Psi \) to estimate \( \sigma^2(\Psi) \):

\[ \widehat{\sigma} = \arg\max_\sigma L_R(\widehat{\Psi}_\sigma, \sigma^2; y) \]

The estimates \( \widehat{\Psi}_\widehat{\sigma} \) and \( \widehat{\sigma} \) maximise \( L_R(\Psi, \sigma^2; y) \).

3. \( \beta \) is then estimated by finding its expected value with respect to the posterior distribution
ML vs REML

- REML makes a less biased estimation of random effects variances than ML.
- The REML estimates of $\Psi$ and $\sigma^2$ are invariant to the value of $\beta$ and less sensitive to outliers in the data compared to ML estimates.
- If you use REML to estimate the parameters, you can only compare two models that have the identical fixed-effects design matrices and are nested in their random-effects terms.
Hierarchical models in R

We could use:

- `nlme` – “Fit and compare Gaussian linear and nonlinear mixed-effects models.”
  
  
  Choose `method = "ML"` or `method = "REML"`

- `lme4` – “Fit linear and generalized linear mixed-effects models.”
  
  More efficient than `nlme`, uses a slightly different syntax, still being developed.

We use `nlme`. (However we will not use `groupedData` objects.)

(If you do try `lme4`, avoid using both `nlme` and `lme4` within the same R session.)
The orthodont example

We consider the model

\[ y_{ij} = \beta_1 + b_j + \beta_2 x_{ij} + \epsilon_{ij} \]

where \( b_j \sim N(0, \sigma_b^2) \) and \( \epsilon_{ij} \sim N(0, \sigma^2) \).

Here

- \( y_{ij} \) is the distance for observation \( i \) on subject \( j \)
- \( x_{ij} \) is the age of the subject when the \( i \)th observation is made on subject \( j \).
- we are regressing distance on age, and allowing a random intercept for each subject (focus on female).

We can fit the model using the \texttt{lme} function as follows:

```r
orthf1.lme <- lme(distance ~ 1 + age,
                   random = ~ 1 | mySubj, data = orthf)
```
> summary(orthf1.lme)

....

Random effects:
  Formula: ~1 | mySubj
       (Intercept) Residual
         StdDev:  2.0685  0.78003

Fixed effects: distance ~ 1 + age

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std.Error</th>
<th>DF</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>17.3727</td>
<td>0.85874</td>
<td>32</td>
<td>20.2304</td>
<td>0</td>
</tr>
<tr>
<td>age</td>
<td>0.4795</td>
<td>0.05259</td>
<td>32</td>
<td>9.1186</td>
<td>0</td>
</tr>
</tbody>
</table>

....

The parameter estimates are:

\[ \hat{\sigma}_b = 2.0685, \quad \hat{\sigma} = 0.78003, \quad \hat{\beta}_1 = 17.3727, \quad \hat{\beta}_2 = 0.4795. \]

The estimated distribution of the intercept is: \( N(\hat{\beta}_1, \hat{\sigma}_b^2) \)
The estimated distribution of the residual is: \( N(0, \hat{\sigma}^2) \)
We obtain the likelihood as follows:

```r
> logLik(orthf1.lme)
 'log Lik.' -70.609 (df=4)
```

For `orthf1.lme`, the estimated fixed effects $\hat{\beta}_1$ and $\hat{\beta}_2$ are

```r
> fixef(orthf1.lme)
 (Intercept)         age
 17.37273   0.47955
```

The average distance as a function of age $x$ is

$$\hat{\beta}_1 + \hat{\beta}_2 x$$
The random effects "estimates" $\hat{b}_j$ are

```r
> ranef(orthf1.lme)
  (Intercept)
  F10   -4.005329
  F06   -1.470449
  ....
  F11    3.599309
```

These "estimates" $\hat{b}_j$ are in fact best linear unbiased predictors (BLUPs) of the random effects, also called the conditional modes (or conditional means). The random effects $b_j$ are actually random variables rather than unknown parameters, and normally we speak of predicting (rather than estimating) random variables.
These "estimates" measure the deviation from the average intercept for each group. The estimated average distance in group \(j\) as a function of age \(x\) is equal to \(\hat{\beta}_1 + \hat{b}_j + \hat{\beta}_2 x\).

To get the coefficients for each subject, i.e. \(\hat{\beta}_1 + \hat{b}_j\) and \(\hat{\beta}_2\), we can use

```r
> coef(orthf1.lme)
    (Intercept)      age
F10   13.367 0.47955
F06   15.902 0.47955
F11   20.972 0.47955

....
```

Again, it is very important to keep in mind that these coefficients are predictions and not estimates.
Tests of random effects

Now consider the model with an additional random effect in the slope:

\[ y_{ij} = \beta_1 + b_{1j} + (\beta_2 + b_{2j})x_{ij} + \epsilon_{ij} \]

where \( y_{ij} \) is the distance for observation \( i \) on subject \( j \).

We fit this model in R as follows:

```r
orthf2.lme <- lme(distance ~ 1 + age,
                  random = ~ 1 + age | mySubj, data = orthf)
```

What are the parameters of this model?

- two fixed-effect \( \beta_1, \beta_2 \)
- the variance of \( \epsilon_{ij} \) denoted by \( \sigma \)
- the variances of \( b_{1j} \) and \( b_{2j} \) denoted respectively by \( \sigma_1^2 \) and \( \sigma_2^2 \)
- and the covariance of \( b_{1j} \) and \( b_{2j} \) denoted by \( \sigma_{12} \).
Tests of random effects

Compared with the more complex model `orthf2.lme`, is the simpler model `orthf1.lme` adequate?

- These two models are nested: `orthf1.lme` is a special case of `orthf2.lme`.
- We can use the Likelihood ratio test to compare nested models.

If $L_2$ is the maximised likelihood of the more general model (e.g. `orthf2.lme`) and $L_1$ is the maximised likelihood of the simpler model (e.g. `orthf1.lme`), then the test statistic is

$$2 \log(L_2/L_1) = 2[\log L_2 - \log L_1].$$

Let $k_i$ be the number of parameters of model $i$ (e.g. $k_1 = 4$, $k_2 = 6$). Then under the null hypothesis that the simpler model is true, the large sample distribution of this test statistic is approximately $\chi^2$ with $k_2 - k_1$ degrees of freedom.
We can perform this test using the `anova` function:

```r
> anova(orthf1.lme, orthf2.lme)

                        df  AIC   BIC logLik Test L.Ratio p-value
orthf1.lme             4 149  156  -70.609
orthf2.lme             6 149  159  -68.714 1 vs 2   3.78  0.1503
```

\[ p\text{-value}=p\left(\chi_{k_2-k_1}^2 \geq 2 \log\left(L_2/L_1\right)\right) \]

So we do not have enough evidence to reject the null hypothesis that the simpler model is adequate.
Conditions to use Likelihood ratio tests (LRT)

- one model must be nested within the other
- the fitting method, i.e. REML or ML, must be the same for both models (and if using REML the fixed parts of the two models must be the same – need to fit using ML otherwise).
- Under the null hypothesis the parameter $\sigma^2$ is set to zero, which is on the boundary of the set of possible values for $\sigma^2$. A consequence is that, in addition to being approximate (since based on an asymptotic result) $p$-values such as the one above tend to be conservative – i.e. the $p$-value above may be larger than the true $p$-value.
- Possible to check on the distribution of the likelihood ratio test under the null hypothesis through simulation.

See Pinheiro and Bates (2000), Section 2.4.1, and Snijders and Bosker (2012), Section 6.2.1.
We can also compare against model with no random effects:

\[ y_{ij} = \beta_1 + b_j + \beta_2 x_{ij} + \epsilon_{ij} \]

where \( b_j \sim N(0, \sigma_b^2) \) and \( \epsilon_{ij} \sim N(0, \sigma^2) \) v.s.

\[ y_{ij} = \beta_1 + \beta_2 x_{ij} + \epsilon_{ij} \].

```r
> orthf.lm <- lm(distance ~ 1 + age, data = orthf)
> orthf1ML.lme <- lme(distance ~ 1 + age,
                  random = ~ 1 | mySubj,
                  data = orthf, method = "ML")
> anova(orthf1ML.lme, orthf.lm)

Model df AIC  BIC logLik Test L.Ratio p-value
orthf1ML.lme 1  4 146.03 153.17 -69.0
orthf.lm 2  3 196.76 202.11 -95.4 1 vs 2   52.725 <.0001
```

The random effects are an important part of our model.
Tests of fixed effects

We can test the effect of a single fixed parameter using an approximate test for which the test statistic is

\[
\frac{\hat{\beta}_k}{\text{std. error}(\hat{\beta}_k)}.
\]

> summary(orthf1.lme)

....
Fixed effects: distance ~ 1 + age

<table>
<thead>
<tr>
<th>Value</th>
<th>Std.Error</th>
<th>DF</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept) 17.3727 0.85874 32 20.2304 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>age 0.4795 0.05259 32 9.1186 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Both fixed-effects terms are clearly significant.
Likelihood ratio tests of fixed effects are also possible – estimation must be via maximum likelihood.

Use full Orthodont dataset from Pinheiro and Bates (2000).

```r
> orth1ML.lme <- lme(distance ~ 1 + age,
                   random = ~ 1 + age | Subject,
                   data = Orthodont, method = "ML")
> orth2ML.lme <- lme(distance ~ 1 + age*Sex,
                   random = ~ 1 + age | Subject,
                   data = Orthodont, method = "ML")
> anova(orth1ML.lme, orth2ML.lme)

          Model df   AIC   BIC logLik Test L.Ratio p-value
orth1ML.lme 1  6 451.21 467.30 -219.61
orth2ML.lme 2  8 443.81 465.26 -213.90 1 vs 2  11.406 0.0033
```

The p-value is small, so we prefer orth2ML.lme. (Pinheiro and Bates (2000), Section 2.4.2, are very cautious about LRTs of this type: they show that such tests can be “anti-conservative”, giving p-values that are too small.)
Confidence intervals

- Confidence intervals on the variance-covariance components and the fixed effects can be obtained using approximate distribution for the maximum likelihood estimates and the RMLE estimates.
- See Pinheiro and Bates (2000), Section 2.4.3.
- Use function `intervals` in R.
Confidence intervals

Recall that $\hat{\beta}_1 = 17.3727$, $\hat{\beta}_2 = 0.4795$, $\hat{\sigma}_b = 2.0685$, $\hat{\sigma} = 0.78003$.

> intervals(orthf1.lme)

Approximate 95% confidence intervals

Fixed effects:

<table>
<thead>
<tr>
<th></th>
<th>lower</th>
<th>est.</th>
<th>upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>15.62353</td>
<td>17.37273</td>
<td>19.12193</td>
</tr>
<tr>
<td>age</td>
<td>0.37242</td>
<td>0.47955</td>
<td>0.58667</td>
</tr>
</tbody>
</table>

attr(,"label")

[1] "Fixed effects:

Random Effects:

Level: mySubj

<table>
<thead>
<tr>
<th></th>
<th>lower</th>
<th>est.</th>
<th>upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>sd((Intercept))</td>
<td>1.3138</td>
<td>2.0685</td>
<td>3.2567</td>
</tr>
</tbody>
</table>

Within-group standard error:

<table>
<thead>
<tr>
<th></th>
<th>lower</th>
<th>est.</th>
<th>upper</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.61052</td>
<td>0.78003</td>
<td>0.99662</td>
</tr>
</tbody>
</table>
Examining residuals

See Pinheiro and Bates (2000), Section 4.3. Our models assume

• within-group errors are i.i.d. and normally distributed, with mean zero and variance $\sigma^2$

• the random effects are normally distributed with mean zero and some covariance matrix.
Residuals and plots

Consider the model

```r
orthf1.lme <- lme(distance ~ 1 + age,
                 random = ~ 1 | mySubj, data = orthf)
```

\[ y_{ij} = \beta_1 + b_j + \beta_2 x_{ij} + \epsilon_{ij} \]

The within-group residuals are defined by

\[ e_{ij} = y_{ij} - (\hat{\beta}_1 + \hat{b}_j + \hat{\beta}_2 x_{ij}) \]

and can be obtained by `resid(orthf1.lme)`. These are unstandardised. To get standardised residuals, use

`resid(orthf1.lme, type = "pearson")`

The type argument can be abbreviated to `type = "p"`. 
```r
# resid(.) means the residuals of orthf1.lme
plot(orthf1.lme, mySubj ~ resid(.), abline = 0, ylab = "Subject")
```
qqnorm(orthf1.lme, ~ resid(. , type = "p"))
Further examples to try (see Pinheiro and Bates (2000), Section 4.3).

\[
\text{orth2.lme} \leftarrow \text{lme(distance} \sim 1 + \text{age*Sex,} \\
\text{random} = \sim 1 + \text{age} | \text{Subject,} \\
\text{data} = \text{Orthodont})
\]

\[
\text{plot(orth2.lme, Subject} \sim \text{resid(.)}, \text{abline} = 0)
\]

\[
\text{plot(orth2.lme, resid(.), type = "p") \sim fitted(.) | Sex,} \\
\text{id} = 0.05, \text{adj} = -0.3
\]

\[
\text{qqnorm(orth2.lme, resid(.), type = "p") | Sex)}
\]

\[
\text{qqnorm(orth2.lme, ranef(.), id} = 0.1
\]

\[
\text{pairs(orth2.lme, ranef(.)) | Sex,} \\
\text{id} = \sim \text{Subject} == "M13", \text{adj} = -0.3
\]

Also see Pinheiro and Bates (2000), Section 4.3, for a model where the within-subject variance differs between Males and Females.
3. Inferring the parameters of a hierarchical linear model: a Bayesian approach
Bayesian models

• In our models so far, the unknown parameters $\beta_1, \beta_2, \sigma^2, \sigma^2_b, \ldots$, have been treated as fixed unknown constants – the classical/frequentist approach.

• Recall that in Bayesian statistics we represent our uncertainty about a parameter $\theta$ (which in general is a vector) by giving $\theta$ a probability distribution (i.e. we treat $\theta$ as a random variable).
Bayesian models

We have

- a model $p(y \mid \theta)$ for the data, conditional on the parameter being $\theta$
- and a prior distribution $p(\theta)$ for the parameter $\theta$.

Then the posterior distribution $p(\theta \mid y)$ for $\theta$ given $y$ is

$$p(\theta \mid y) \propto p(y \mid \theta)p(\theta)$$

posterior $\propto$ likelihood $\times$ prior.
The form of a Bayesian hierarchical model is as follows.

We have

- a model $p(y_{ij} | \theta_j)$ for the data in group $j$ (for $j = 1, \ldots, J$)
- a between-group model or “prior” $p(\theta_j | \phi)$ for the parameter $\theta_j$ (for $j = 1, \ldots, J$)
- and a prior $p(\phi)$ for the hyperparameter $\phi$. 
Example: schools

- Consider an educational study with data from students in many schools
- We want to predict the students’ grades $y$ on a standardized test given their scores on a pre-test $x$

$$y \sim \mathcal{N}(b_1 + b_2 x, \sigma^2)$$

- The regression model can depend on the school

$$y_{ij} \sim \mathcal{N}(b_{j1} + b_{j2} x_{ij}, \sigma^2)$$

where

$$
\begin{pmatrix}
  b_{j1} \\
  b_{j2}
\end{pmatrix}
\sim
\mathcal{N}
\left(
\begin{pmatrix}
  \beta_1 \\
  \beta_2
\end{pmatrix},
\Psi
\right)
$$
Example: schools

We consider \( y_{ij} \sim N(b_{j1} + b_{j2}x_{ij}, \sigma^2) \) where

\[
\begin{pmatrix}
b_{j1} \\
b_{j2}
\end{pmatrix} \sim N\left(\begin{pmatrix}
\beta_1 \\
\beta_2
\end{pmatrix}, \Psi\right).
\]
Bayesian parameter inference

• In Bayesian statistics, the information about a parameter $\phi$ is summarised in the posterior distribution.
• A parameter value with a high posterior probability explains more the data than a parameter value with a lower posterior probability.
• The posterior probability distribution is typically not available in closed form.
• There is therefore a need to use computer simulation and Monte-Carlo techniques to sample from the posterior distribution.
Markov Chain Monte-Carlo algorithm

**Idea:** construct a Markov chain with stationary distribution \( p(\phi|x^*) \)

**Metropolis-Hasting algorithm:**
- start from an arbitrary point \( \phi^{(0)} \)
- for \( t = 1, \ldots \)
  - generate \( \tilde{\phi} \sim q(\tilde{\phi}|\phi^{(t)}) \)
  - take \( \phi^{(t+1)} = \begin{cases} \tilde{\phi} & \text{with probability } p = \min \left( 1, \frac{p(y|\tilde{\phi})p(\tilde{\phi})q(\phi^{(t)}|\tilde{\phi})}{p(y|\phi^{(t)})p(\phi^{(t)})q(\phi^{(t)}|\phi^{(t)})} \right) \\ \phi^{(t)} & \text{with probability } 1 - p \end{cases} \)
MCMC and Gibbs Sampler

- We generate samples from the posterior distribution of the model parameters using MCMC (Markov Chain Monte Carlo) simulation. Much more on MCMC in Further Statistical Methods.
- To do so, we need to be able to evaluate the likelihood $p(y|\phi)$ for each value of $\phi$.
- However, this require solving the following integral:

$$p(y|\phi) = \int p(y|\theta)p(\theta|\phi)d\theta$$

- The Gibbs Sampler algorithm can be used in such a case to approximate the posterior distribution.
In R: rjags, JAGS

• We use rjags, which depends on coda. Installing rjags will normally install coda as well – use the Packages menu, or alternatively something like `install.packages("rjags")`
• You also need to install JAGS (Just Another Gibbs Sampler) which you can get from http://sourceforge.net/projects/mcmc-jags/files/JAGS/3.x/
Radon data

Example from Gelman and Hill (2007).

- The entire dataset consists of radon measurements from over 80,000 houses spread across approximately 3,000 counties in the USA. As in Gelman and Hill (2007), Chapter 12, we will restrict to estimating the distribution of radon levels within the 85 counties in Minnesota (919 measurements).

- Think of the measurements made in the same county as a group.

```r
### radon_setup.R from
### http://www.stat.columbia.edu/~gelman/arm/examples/radon/
## a copy is available on the HMs webpage
source("radon_setup.R")
```
```r
> radondata <- data.frame(y = y, x = x, county = county)
> radondata
         y   x county
 1 0.788461 1  1
 2 0.788461 0  1
 3 1.064711 0  1
 4 0.000000 0  1
 5 1.131402 0  2
 917 1.609440 0 84
 918 1.308330 0 85
 919 1.064710 0 85

• *y* is the log of the radon measurement in a house
• *x* is the floor of this measurement (*x* = 0 for basement, *x* = 1 for first floor)
• *county* is the county in which the measurement was taken.

At first we ignore *x*.
Models

Let the observations be indexed by $i = 1, \ldots, n$ (where $n = 919$) and let the counties be indexed by $j = 1, \ldots, J$ (where $J = 85$). Let $j[i]$ be shorthand for county $[i]$, so


Consider the model

$$y_i = \beta_j[i] + \epsilon_i$$

where $\beta_j \sim N(\mu, \sigma^2_b)$ and $\epsilon_i \sim N(0, \sigma^2)$. This is the same model as for the Rail data, just written in a slightly different way.
We have

- \( y_i \sim N(\beta_{j[i]}, \sigma^2) \)
- where \( \beta_j \sim N(\mu, \sigma^2_b) \)
- and where \( \mu, \sigma^2 \) and \( \sigma^2_b \) are fixed unknown parameters.

Following a Bayesian approach:

- We treat the unknown parameters as random variables.
- After defining the likelihood and the priors for the parameters, we use an MCMC algorithm (via JAGS) to sample from the appropriate posterior distribution.
- MCMC algorithms generate dependent samples from the posterior distribution, using the likelihood and priors.
The first step is to tell JAGS what the model is, using a model file. We create a text file called radon1.jags containing the following.

```jags
model {
    for(i in 1:n) {
        y[i] ~ dnorm(beta[county[i]], sigma^(-2))
    }
    for(j in 1:J) {
        beta[j] ~ dnorm(mu, sigma.b^(-2))
    }
    mu ~ dnorm(0, 0.0001)
    sigma ~ dunif(0, 100)
    sigma.b ~ dunif(0, 100)
}
```
• Note: JAGS uses the *precision*, not the variance, when specifying a normal distribution (precision = 1/variance). So the precision of $y_i$ being $\sigma^{-2}$ corresponds to a variance of $\sigma^2$.

• The prior for $\mu$ is a normal with mean 0 and precision 0.0001, i.e. a variance of 1000. The priors for $\sigma$ and $\sigma_b$ are both $\text{Uniform}(0, 100)$.

• R needs to be able to find radon1.jags. Make sure you start R in a folder where you put this file.
Continue setup.

```r
library(rjags)

r1.inits <- list(beta = rnorm(J), mu = rnorm(1),
                 sigma = runif(1), sigma.b = runif(1))

r1.jags <- jags.model("radon1.jags", inits = r1.inits)
%## the following also works
%## r1.jags <- jags.model("radon1.jags")

Specify the variables to monitor and then run.

r1.vars <- c("beta", "mu", "sigma", "sigma.b")

r1.sim <- coda.samples(r1.jags, r1.vars, n.iter = 10000)
```
> summary(r1.sim)

Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable, plus standard error of the mean:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
<th>Naive SE</th>
<th>Time-series SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta[1]</td>
<td>1.0595</td>
<td>0.2525</td>
<td>0.002525</td>
<td>0.0022662</td>
</tr>
<tr>
<td>beta[2]</td>
<td>0.8891</td>
<td>0.1068</td>
<td>0.001068</td>
<td>0.0012457</td>
</tr>
<tr>
<td>beta[3]</td>
<td>1.2298</td>
<td>0.2648</td>
<td>0.002648</td>
<td>0.0025880</td>
</tr>
<tr>
<td>....</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>beta[85]</td>
<td>1.2828</td>
<td>0.2773</td>
<td>0.002773</td>
<td>0.0027516</td>
</tr>
<tr>
<td>mu</td>
<td>1.3118</td>
<td>0.0499</td>
<td>0.000499</td>
<td>0.0007692</td>
</tr>
<tr>
<td>sigma</td>
<td>0.7994</td>
<td>0.0195</td>
<td>0.000195</td>
<td>0.0003554</td>
</tr>
<tr>
<td>sigma.b</td>
<td>0.3162</td>
<td>0.0490</td>
<td>0.000490</td>
<td>0.0016397</td>
</tr>
</tbody>
</table>
2. Quantiles for each variable:

<table>
<thead>
<tr>
<th></th>
<th>2.5%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta[1]</td>
<td>0.5471</td>
<td>0.8922</td>
<td>1.0638</td>
<td>1.2306</td>
<td>1.5466</td>
</tr>
<tr>
<td>beta[2]</td>
<td>0.6813</td>
<td>0.8179</td>
<td>0.8890</td>
<td>0.9596</td>
<td>1.0981</td>
</tr>
<tr>
<td>beta[3]</td>
<td>0.6985</td>
<td>1.0596</td>
<td>1.2308</td>
<td>1.4071</td>
<td>1.7483</td>
</tr>
<tr>
<td>beta[85]</td>
<td>0.7448</td>
<td>1.0995</td>
<td>1.2792</td>
<td>1.4662</td>
<td>1.8331</td>
</tr>
<tr>
<td>mu</td>
<td>1.2137</td>
<td>1.2783</td>
<td>1.3119</td>
<td>1.3446</td>
<td>1.4117</td>
</tr>
<tr>
<td>sigma</td>
<td>0.7618</td>
<td>0.7860</td>
<td>0.7992</td>
<td>0.8123</td>
<td>0.8388</td>
</tr>
<tr>
<td>sigma.b</td>
<td>0.2245</td>
<td>0.2817</td>
<td>0.3149</td>
<td>0.3487</td>
<td>0.4146</td>
</tr>
</tbody>
</table>

Using `summary(r1.sim[, c(1:3, 85:88)])` picks out just the parameters we have looked at above. These results are similar to those obtained from the classical approach using

```r
r1.lme <- lme(y ~ 1, random = ~ 1 | county)
```
In these trace plots we are looking to see if the Markov chain has mixed well, i.e. to see if it appears that it has (approximately) reached its stationary distribution.

\texttt{xyplot(r1.sim[, c(1, 86:88)])}
densityplot(r1.sim[, c(1:3, 84:88)])
95% HPD intervals.

```r
> HPDinterval(r1.sim)
[[1]]

<table>
<thead>
<tr>
<th></th>
<th>lower</th>
<th>upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta[1]</td>
<td>0.5712122</td>
<td>1.5625728</td>
</tr>
<tr>
<td>beta[2]</td>
<td>0.6789466</td>
<td>1.0955027</td>
</tr>
<tr>
<td>beta[3]</td>
<td>0.6786186</td>
<td>1.7246411</td>
</tr>
<tr>
<td>....</td>
<td></td>
<td></td>
</tr>
<tr>
<td>beta[84]</td>
<td>1.1372023</td>
<td>1.8465939</td>
</tr>
<tr>
<td>beta[85]</td>
<td>0.7216640</td>
<td>1.8096207</td>
</tr>
<tr>
<td>mu</td>
<td>1.2128849</td>
<td>1.4098202</td>
</tr>
<tr>
<td>sigma</td>
<td>0.7595627</td>
<td>0.8359080</td>
</tr>
<tr>
<td>sigma.b</td>
<td>0.2271229</td>
<td>0.4163161</td>
</tr>
<tr>
<td>....</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```
Suppose we now want to incorporate the effect of the floor $x$ on which the measurement was taken ($x = 0$ for basement, $x = 1$ for first floor).

We make our Bayesian model

- $y_i \sim N(\beta_j[i] + \alpha x_i, \sigma^2)$
- where $\beta_j \sim N(\mu, \sigma^2_b)$
- and where $\mu$, $\alpha$, $\sigma^2$ and $\sigma^2_b$ have prior distributions.

[Our notation differs from that of Gelman and Hill (2007): where we have $\beta_j$ they use $\alpha_j$, and where we have $\alpha$ they use $\beta$.]
We create a second file called radon2.jags as follows.

```jags
model {
  for(i in 1:n) {
    y[i] ~ dnorm(beta[county[i]] + alpha*x[i], sigma^(-2))
  }
  for(j in 1:J) {
    beta[j] ~ dnorm(mu, sigma.b^(-2))
  }
  mu ~ dnorm(0, 0.0001)
  alpha ~ dnorm (0, 0.0001)
  sigma ~ dunif(0, 100)
  sigma.b ~ dunif(0, 100)
}
```
r2.inits <-
    list(beta = rnorm(J), alpha = rnorm(1), mu = rnorm(1),
         sigma = runif(1), sigma.b = runif(1))

r2.jags <- jags.model("radon2.jags", inits = r2.inits)

r2.vars <- c("beta", "alpha", "mu", "sigma", "sigma.b")

r2.sim <- coda.samples(r2.jags, r2.vars, n.iter = 10000)
> summary(r2.sim[, c(1:3, 86:89)])

Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable, plus standard error of the mean:

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Naive SE</th>
<th>Time-series SE</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>alpha</td>
<td>-0.6932</td>
<td>0.0716</td>
<td>0.0007</td>
<td>0.0011</td>
<td>0.0011</td>
</tr>
<tr>
<td>beta[1]</td>
<td>1.1869</td>
<td>0.2558</td>
<td>0.0026</td>
<td>0.0033</td>
<td>0.0033</td>
</tr>
<tr>
<td>beta[2]</td>
<td>0.9297</td>
<td>0.1012</td>
<td>0.0010</td>
<td>0.0012</td>
<td>0.0012</td>
</tr>
<tr>
<td>beta[85]</td>
<td>1.3823</td>
<td>0.2836</td>
<td>0.0028</td>
<td>0.0029</td>
<td>0.0029</td>
</tr>
<tr>
<td>mu</td>
<td>1.4607</td>
<td>0.0535</td>
<td>0.0005</td>
<td>0.0009</td>
<td>0.0009</td>
</tr>
<tr>
<td>sigma</td>
<td>0.7568</td>
<td>0.0182</td>
<td>0.0002</td>
<td>0.0003</td>
<td>0.0003</td>
</tr>
<tr>
<td>sigma.b</td>
<td>0.3358</td>
<td>0.0469</td>
<td>0.0004</td>
<td>0.0013</td>
<td>0.0013</td>
</tr>
</tbody>
</table>

....
We can compare the Bayesian results above with the following classical model.

```r
> r2.lme <- lme(y ~ 1 + x, random = ~ 1 | county)
> summary(r2.lme)
....
```

Random effects:

```
  Formula: ~1 | county
    (Intercept)   Residual
  StdDev:  0.3282224  0.7555892
```

Fixed effects: y ~ 1 + x

```
  Value     Std.Error    DF      t-value  p-value
(Intercept) 1.4615979  0.05157622 833  28.338600  0
x            -0.6929937  0.07043081 833  -9.839354  0
....
• The estimates and standard errors for the classical and Bayesian models are similar: for both models the coefficient of \( x \) is estimated as \(-0.7\). Further, a 95% confidence interval, and also a 95% credible interval, are both estimated as approximately \((-0.83, -0.55)\).

• That is, the floor on which the radon measurement is made has a significant effect on the measurement: in comparison with basement measurements, a measurement on the first floor is lower (on average, lower by 0.7).
Gelman and Hill (2007) avoid the term “random effects.” They prefer the term “modelled” for coefficients that are modelled using a probability distribution (i.e. what we have been calling random effects), and “unmodelled” otherwise (i.e. what we have been calling fixed effects). Their recommended strategy for multilevel modelling is:

• start by fitting classical regressions (e.g. using `lm`), . . .
• then setup multilevel models and fit using `lme`, . . .
• then fit fully Bayesian multilevel models, . . .
• . . .

Getting as far as fitting Bayesian models means treating all coefficients as being “modelled.”
4. Studying inferred Hierarchical Linear Models and cross-level interactions

- Between and within group variability
- Prediction
- Cross-level interactions
Hierarchical Linear Models

A varying-intercept, varying-slope linear model (with one explanatory variable) can be written

\[ y_{ij} = \beta_1 + b_j + (\beta_2 + c_j) x_{ij} + \epsilon_{ij} \]

where

- \( y_{ij} \) is the \( i \)-th observation in the \( j \)-th group
- \( x_{ij} \) is the \( i \)-th explanatory variable in the \( j \)-th group
- \( \beta_1, \beta_2 \) are fixed effects
- \( (b_j, c_j) \sim N(0, \Psi) \) are the random effects
- \( \epsilon_{ij} \sim N(0, \sigma^2) \)
- \( (b_j, c_j) \) and \( \epsilon_{ij} \) are mutually independent and independent across groups
More general formula

Denote by:
- $y_j$ the vector of observations in group $j$
- $\beta$ the vector of fixed parameters (with components $\beta_1, \beta_2, \ldots$)
- $b_j$ the vector of random effects for group $j$ (with components $b_{1j}, b_{2j}, \ldots$)
- $\epsilon_j$ the vector of random errors for group $j$.

Then, any hierarchical linear model (with one level of grouping) can be written as

$$y_j = X_j\beta + Z_j b_j + \epsilon_j$$

where the matrices $X_j$ and $Z_j$ are the known fixed-effects and random-effects regressor matrices.
• The components of the vector $\epsilon_j$ are i.i.d.

\[ \epsilon_j \sim \mathcal{N}(0, \sigma^2 I) . \]

• The vectors $b_j$ are assumed to be independent and normally distributed with mean 0:

\[ b_j \sim \mathcal{N}(0, \Psi) \]

• $b_j$ and $\epsilon_j$ are mutually independent and independent across groups

Both frequentist and Bayesian approaches enable us to
• infer the parameters of the model
• obtain confidence interval/regions regarding each parameter
Example: language scores in elementary schools

Example from Snijders and Bosker (2012).

Data on 3758 students, age about 11 years, from 211 schools in the Netherlands. The structure of the data is students within classes, where class sizes range from 4 to 34. The dataset contains one class per school, so the class and school level are the same.

Let $y$ be a student’s score on a language test, and let IQ be a measure of their verbal IQ. We consider the model

$$y_{ij} = \beta_0 + \beta_1 \text{IQ}_{ij} + b_j + \epsilon_{ij}$$

where $j$ indexes schools, and $i$ indexes the students within a school. Here $\beta_0, \beta_1$ are fixed parameters and, independently, $b_j \sim N(0, \sigma^2_b)$ and $\epsilon_{ij} \sim N(0, \sigma^2)$. The IQ score has been centered so that its mean is 0 (approx).
Centering the data

• Important to centre the explanatory variables
• If we are interested in the deviation to the mean it may be interesting to add the mean as a variable
• This next point will be covered in the last lecture as well as its link to the between and within group regression coefficients.
In R

```r
> nlschools <- read.table("mlbook2_r.dat", header = TRUE)
> names(nlschools)[c(3, 4, 5, 9, 10)] <-
    c("langscore", "SES", "IQ", "SESsch.av", "IQsch.av")
> lang.lme <- lme(langscore ~ 1 + IQ, random = ~ 1 | schoolnr,
    data = nlschools, method = "ML")

> summary(lang.lme)
....
Random effects:
  Formula: ~1 | schoolnr
    (Intercept) Residual
        StdDev: 3.1377 6.3615

Fixed effects: langscore ~ 1 + IQ
  Value Std.Error   DF  t-value  p-value
(Intercept) 41.055    0.243451 3546 168.637      0
     IQ       2.507    0.054396 3546  46.096      0
....
```
• The fitted regression line for school $j$ is

$$y_{ij} = 41.06 + b_j + 2.507IQ_{ij}.$$ 

• $b_j \sim \mathcal{N}(0, \hat{\sigma}_b^2)$ are the school-dependent deviations, with $\hat{\sigma}_b = 3.14$.

• The scatter around the regression line for a particular student has standard deviation $\hat{\sigma} = 6.36$. 
Variance and covariance

The covariance between two students in the same class $j$ is

$$Cov(y_{ij}, y_{ij'}) = \hat{\sigma}_b^2 = 3.14^2$$

and the variance of the score for one student is

$$Var(y_{ij}) = \hat{\sigma}_b^2 + \hat{\sigma}^2 = 3.14^2 + 6.36^2$$

In this varying intercept, fixed-slope model, the covariance and variance do not depend on the IQ.
Covariance for varying-intercept, varying-slope model

Consider a varying-intercept, varying-slope model:

\[ y_{ij} = \beta_1 + b_j + (\beta_2 + c_j)x_{ij} + \epsilon_{ij}. \]

The covariance between two individuals in the same group \( j \) is

\[
\text{Cov}(y_{ij}, y_{i'j}) = \text{Cov}(\beta_1 + b_j + (\beta_2 + c_j)x_{ij} + \epsilon_{ij}, \beta_1 + b_j + (\beta_2 + c_j)x_{i'j} + \epsilon_{i'j})
\]
\[
= \text{Cov}(b_j + c_jx_{ij} + \epsilon_{ij}, b_j + c_jx_{i'j} + \epsilon_{i'j})
\]
\[
= \Psi_{11}^2 + \Psi_{12}(x_{ij} + x_{i'j}) + \Psi_{22}^2x_{ij}x_{i'j}
\]

and the variance of one individual is

\[
\text{Var}(y_{ij}) = \Psi_{11}^2 + 2\Psi_{12}x_{ij} + \Psi_{22}^2x_{ij}^2 + \sigma^2.
\]

Exercise: derive the covariance between two individuals in the same group for a model with more than one explanatory variable.
Prediction in Hierarchical Models

Predictions in hierarchical models is more complicated than in classical regression as we have to decide

• do we want the predicted value for an existing group?
• or do we want the fitted/predicted value for a new group?

In the school example:

• do we want to predict the score of a new student in an existing school?
• or do we want to predict the score of a new student in a new school?
Prediction in classical regression

In classical regression, e.g., linear regression

\[ Y = X\beta + \epsilon. \]

Prediction is simple:

1. Specify the predictor matrix \( \tilde{X} \) for a new set of observations
2. Compute the linear predictor \( \tilde{X}\beta \)
3. Simulate the predictive data by simulating independent normal errors \( \tilde{\epsilon} \) with mean 0 and standard deviation \( \sigma \) and compute \( \tilde{y} = \tilde{X}\beta + \tilde{\epsilon} \)
Prediction for a new observation in an existing group

Consider a varying-intercept, varying-slope model:

\[ y_{ij} = \beta_1 + b_j + (\beta_2 + c_j)x_{ij} + \epsilon_{ij} . \]

If we want to predict the value for a new individual in an existing group, we need to take into account the estimated (or predicted) values \( \hat{b}_j \) and \( \hat{c}_j \) of the random effects.

Let \( \tilde{x} \) the explanatory variable for the new individual in group \( j \), then the value for this new individual can be simulated from a normal distribution as follows

\[ \tilde{y} \sim \mathcal{N}(\hat{\beta}_1 + \hat{b}_j + (\hat{\beta}_2 + \hat{c}_j)\tilde{x}, \sigma^2) \]
Prediction for a new observation in a new group

Consider a varying-intercept, varying-slope model:

\[ y_{ij} = \beta_1 + b_j + (\beta_2 + c_j)x_{ij} + \epsilon_{ij}. \]

If we want to predict the value for a new individual in a new group, the random effect should be set equal to zero.

Let \( \tilde{x} \) the explanatory variable for the new individual in a new group, then the value for this new individual can be simulated from a normal distribution as follows

\[ \tilde{y} \sim \mathcal{N}(\hat{\beta}_1 + \hat{\beta}_2 \tilde{x}, \sigma^2) \]
Orthodont: simple linear growth curves

This is an example of *growth curve* data, or *repeated measures* or *longitudinal* or *panel* data.

Here we think of the measurements made on the same subject as a group.

For each individual $j$,

- $x_{ij}$ is the age of the child at time $i$.
- $y_{ij}$ is the measurement of the distance from the pituitary gland to the pterygomaxillary fissure of the child at time $i$.

We consider a varying-intercept fixed-slope model

$$y_{ij} = \beta_1 + b_j + \beta_2 x_{ij}$$
Fitted values for an existing group: the random effects should take their estimated values, so we want \( \hat{\beta}_1 + \hat{b}_j + \hat{\beta}_2 x_{ij} \), obtained by

\[
> \text{fitted(orthf1.lme, level = 1)}
\]

<table>
<thead>
<tr>
<th>F01</th>
<th>F01</th>
<th>F01</th>
<th>F01</th>
<th>F02</th>
<th>F02</th>
<th>F02</th>
<th>F02</th>
<th>F02</th>
</tr>
</thead>
</table>

Fitted/predicted values for a new group: the random effects should be set to zero, so we want \( \hat{\beta}_1 + \hat{\beta}_2 x_{ij} \), obtained by

\[
> \text{fitted(orthf1.lme, level = 0)}
\]

<table>
<thead>
<tr>
<th>F01</th>
<th>F01</th>
<th>F01</th>
<th>F01</th>
<th>F02</th>
<th>F02</th>
<th>F02</th>
<th>F02</th>
<th>F02</th>
</tr>
</thead>
</table>

These are for the ages at which we observed F01, F02, ....

\[
\text{fitted(orthf1.lme, level = 0:1)} \quad \text{gives both types.}
\]
Predicted values at other ages:

```r
> neworthf <-
    data.frame(mySubj = rep(c("F02", "F10"), each = 3),
                age = rep(16:18, 2))
> neworthf
   mySubj age
 1    F02  16
 2    F02  17
 3    F02  18
 4    F10  16
 5    F10  17
 6    F10  18

> predict(orthf1.lme, newdata = neworthf, level = 0:1)
   mySubj predict.fixed predict.mySubj
 1    F02     25.045       25.386
 2    F02     25.525       25.865
 3    F02     26.005       26.345
 4    F10     25.045       21.040
 5    F10     25.525       21.520
 6    F10     26.005       21.999
```
Cross-level interaction effects

It is possible to add explanatory variables for each group. For example, we can consider a model of the form

$$ y_{ij} = \gamma_0 j + \gamma_1 j x_{ij} + \epsilon_{ij} $$

where

$$ \gamma_0 j = \beta_{00} + \beta_{01} z_j + b_{0j} $$

$$ \gamma_1 j = \beta_{10} + \beta_{11} z_j + b_{1j} $$.

Here $x_{ij}$ is an individual-level explanatory variable, and $z_j$ is a group-level explanatory variable. Substituting, we get

$$ y_{ij} = \beta_{00} + \beta_{01} z_j + \beta_{10} x_{ij} + \beta_{11} z_j x_{ij} $$

$$ + b_{0j} + b_{1j} x_{ij} + \epsilon_{ij}. $$

The term $\beta_{11} z_j x_{ij}$ is called the cross-level interaction effect.
Cross-level interaction effects

In this example

\[ y_{ij} = \beta_{00} + \beta_{01} z_j + \beta_{10} x_{ij} + \beta_{11} z_j x_{ij} + b_{0j} + b_{1j} x_{ij} + \epsilon_{ij}. \]

we have fixed-effects for \( x, z \) and their interaction, and (within each group) a random intercept and a random slope for \( x \).

Other combinations of fixed and random effects are possible. For example, we could have a random effect for \( z \) and for the interaction between \( x \) and \( z \).
Cross-level interaction effects: Orthodont example

So far we only focused on female patients.

We want to include the sex as an additional variable: for each individual $j$, $z_j$ denotes the gender of the child (male=0, female=1).

We extend the previous varying-intercept fixed-slope model

$$y_{ij} = \beta_1 + b_j + \beta_2 x_{ij}$$

by making the intercept and the slope depending on the sex of the individual and adding a random effect in the slope. We have:

$$y_{ij} = \beta_1 + b_{1j} + \beta_2 z_i + (\beta_3 + b_{2j} + \beta_4 z_j)x_{ij} + \epsilon_{ij}$$
Given this model

\[ y_{ij} = \beta_1 + b_{1j} + \beta_2 z_i + (\beta_3 + b_{2j} + \beta_4 z_j)x_{ij} + \epsilon_{ij} \]

we can make the following comments.

- The average distance for a boy of age \( a \) is \( \beta_1 + \beta_3 a \).
- The average distance for a girl of age \( a \) is \( \beta_1 + \beta_2 + (\beta_3 + \beta_4) a \).
- The growth rate for girl \( \sim \mathcal{N}(\beta_3 + \beta_4, \sigma_2^2) \).
- The distance distribution for a girl at age \( a \) is a normal with mean \( \beta_1 + \beta_2 + (\beta_3 + \beta_4) a \) and variance \( \sigma_1^2 + 2\sigma_{12} a + a^2 \sigma_2^2 + \sigma^2 \).
A made-up medical study example

In a medical study, each patient receives one of two possible treatments and undergoes a laboratory test every month for a year.

Consider a mixed-effects model in which the outcome of the laboratory test – described by a real number – is modelled as a linear function of the duration of the treatment.

It is assumed that the regression coefficients can differ between patients, and that the slope can depend upon the received treatment.
Let

- $y_{ij} =$ Laboratory test outcome when observation $i$ is made on patient $j$
- $x_{ij} =$ Duration after treatment (in months) when observation $i$ is made on patient $j$
- $z_j =$ Treatment received by patient $j$

for $i = 1, \ldots, 12$ and $j = 1, \ldots, 50$.

Consider the multilevel model:

$$y_{ij} = \beta_1 + b_{1j} + (\beta_2 + \beta_3 z_j + b_{2j})x_{ij} + \epsilon_{ij}$$

where

- $\beta_1, \beta_2, \beta_3$ are fixed effects coefficients,
- $b_{ij}, b_{2,j}$ are random effects for patient $j$,
- and $\epsilon_{ij}$ is the random error for the observation $i$ on patient $j$. 
The distributional assumptions are the following:

- $b_{1j}, b_{2j}$ and $\epsilon_{ij}$ are mutually independent and independent across $j$
- $\epsilon_{ij}$ follows a normal distribution with mean 0 and variance $\sigma^2$,
- $(b_{1j}, b_{2j})'$ follows a bivariate normal distribution with mean $(0, 0)'$ and covariance matrix

$$
\begin{pmatrix}
\sigma_1^2 & \sigma_{12} \\
\sigma_{12} & \sigma_2^2
\end{pmatrix}
$$
What is the variance of the outcome of the laboratory test after \( t \) months of treatments as a function of the parameters?

For a patient \( j \), the variance of the outcome of the test after \( t \) months of treatments is

\[
\text{Var}(y_{ij} | x_{ij} = t) = \sigma_1^2 + 2\sigma_{12} t + \sigma_2^2 t^2 + \sigma^2 .
\]

What is the correlation between the outcomes of the laboratory test of a patient after \( t_1 \) and after \( t_2 \) months of treatment?

The covariance between the outcome of the test after \( t_1 \) and \( t_2 \) months of treatment is:

\[
\text{Cov}(y_{ij}, y_{i'j} | x_{ij} = t_1, x_{i'j} = t_2) = \sigma_1^2 + \sigma_{12}(t_1 + t_2) + \sigma_2^2 t_1 t_2 .
\]

Therefore, the correlation of the outcome of the test after \( t_1 \) and \( t_2 \) months of treatment for a given patient is:

\[
\frac{\sigma_1^2 + \sigma_{12}(t_1 + t_2) + \sigma_2^2 t_1 t_2}{\sqrt{(\sigma_1^2 + 2\sigma_{12} t_1 + \sigma_2^2 t_1^2 + \sigma^2)(\sigma_1^2 + 2\sigma_{12} t_2 + \sigma_2^2 t_2^2 + \sigma^2)}}
\]
The following table gives the results of the model fitted to the data. The variable Duration is the treatment duration measured in months, and the variable Treatment is 0 (for the first treatment) or 1 (for the second treatment).

<table>
<thead>
<tr>
<th>Fixed effects</th>
<th>Value</th>
<th>StdError</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>6.32</td>
<td>0.31</td>
</tr>
<tr>
<td>Duration</td>
<td>4.23</td>
<td>0.33</td>
</tr>
<tr>
<td>Treatment x Duration</td>
<td>1.74</td>
<td>0.46</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random effects</th>
<th>Std.Dev.</th>
<th>Correl.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.82</td>
<td>(Intr)</td>
</tr>
<tr>
<td>Duration</td>
<td>1.63</td>
<td>0.21</td>
</tr>
<tr>
<td>Residual</td>
<td>2.05</td>
<td></td>
</tr>
</tbody>
</table>
What is the estimated average laboratory test outcome after a year of treatment for patients having received the first (resp. second) treatment?

Estimated average test outcome after a year of treatment for patients having received treatment 1:

$$\hat{\beta}_1 + 12\hat{\beta}_2 = 57.08$$

Estimated average test outcome after a year of treatment for patients having received treatment 2:

$$\hat{\beta}_1 + 12(\hat{\beta}_2 + \hat{\beta}_3) = 77.96$$
What is the estimated distribution of growth of the outcome of the laboratory test per month for each treatment?

Growth of the outcome per month for treatment $z$ follows a normal distribution with mean $\hat{\beta}_2 + \hat{\beta}_3 z$ and variance $\sigma^2_z$.

For treatment 1, the mean of the normal distribution is 4.23 whereas for treatment 2 it is 5.97; for both cases the variance is 2.66.
What is the correlation between the outcomes of the laboratory test after 6 and after 12 months of treatment?

The estimated covariance of the outcome of the test after 6 and 12 months of treatment is

\[
Cov(y_{ij}, y_{i'j} | x_{ij} = 6, x_{i'j} = 12) = \hat{\sigma}_1^2 + 18\hat{\sigma}_{12} + 72\hat{\sigma}_2^2 = 198.39
\]

In addition the variance of the outcome is

\[
Var(y_{ij} | x_{ij}) = \hat{\sigma}_1^2 + 2\hat{\sigma}_{12}x_{ij} + \hat{\sigma}_2^2x_{ij}^2 + \sigma^2
\]

which is equal to 105.68 for \(x_{ij} = 6\) and to 395.15 for \(x_{ij} = 12\).

Therefore the correlation of the outcome of the test after 6 and 12 months of treatment is

\[
\frac{Cov(y_{ij}, y_{i'j} | x_{ij} = 6, x_{i'j} = 12)}{\sqrt{Var(y_{ij} | x_{ij} = 6)Var(y_{i'j} | x_{i'j} = 12)}} = \frac{198.39}{\sqrt{105.68 * 395.15}} = 0.97
\]
4. Summary and Further Properties of Hierarchical Models
Motivations for multi-level modelling

Assume you have a dataset that can be separated according to groups.

Hierarchical models or multi-level models can be useful for:

• Learning about treatment effects that vary
• Using all the data to perform inferences for groups with small sample size
• Prediction when data vary with groups
• Analysis of structured data
• More efficient inference for regression parameters
• Including predictors at two different levels
• Getting the right standard error: accurately accounting for uncertainty in prediction and estimation
Fixed effects or random effects?

When should we incorporate random-effects??

See Snijders and Bosker (2012), Section 4.3.1.

- If groups are unique entities and inference should focus on these groups, then *Fixed*. This is often the case with a small number of groups.
- If groups are regarded as sample from a (perhaps hypothetical) population and inference should focus on this population, then *Random*. This often is the case with a large number of groups.
- If group-level effects are to be tested, then *Random*. Reason: a fixed-effects model already explains all differences between groups by the fixed (group) effects, so there is no unexplained between-group variability left that could be explained by group-level variables.
• If group sizes are small and there are many groups (and the assumptions of the model are reasonable), then Random makes better use of the data.

• If the researcher is interested only in within-group effects, and is suspicious about the model for between-group differences, then Fixed is more robust.

• If the assumption that group effects, etc., are normally distributed is a very poor approximation, then Random may lead to unreliable results.

Hierarchical modelling as partial pooling

Gelman and Hill (2007) and Jackman (2009) discuss hierarchical models as being a compromise between two alternative models:

- a *no-pooling* model, and
- a *complete-pooling* model.

Hierarchical models can be thought of as giving *semi-pooling* or *partial-pooling* estimates that lie somewhere between the no-pooling and complete-pooling alternatives.
Radon data

Example from Gelman and Hill (2007). The radon measurements in county $j$

$$y_{ij} \sim \mathcal{N}(\alpha_j + \beta x_{ij}, \sigma^2)$$

where

$$\alpha_j \sim \mathcal{N}(\mu, \sigma_b^2).$$

The parameters $\mu$, $\sigma^2$, $\sigma_b^2$ are unknown.

- We want to estimate the average log radon level $\alpha_j$ for county $j$ (and the other parameters).
- We have $n_j$ observations available in county $j$ (where $n_1 = 4, \ldots, n_{85} = 2$).
No pooling – one explanatory variable

Classical:
r1.unpooled <- lm (formula = y ~ - 1 + x + factor(county))

Bayesian: \( y_i \sim N(\alpha_j[i] + \beta x_i, \sigma_y^2) \) with priors for the \( \alpha_j, \beta \) and \( \sigma_y \)

model {
    for (i in 1:n){
        y[i] ~ dnorm (y.hat[i], tau.y)
        y.hat[i] <- a[county[i]] + b*x[i]
    }
    b ~ dnorm (0, .0001)
    tau.y <- pow(sigma.y, -2)
    sigma.y ~ dunif (0, 100)

    for (j in 1:J){
        a[j] ~ dnorm (0, .0001)
    }
}
Complete pooling – one explanatory variable

Classical: `r1.pooled <- lm(y ~ x)`

Bayesian: \( y_i \sim N(\alpha + \beta x_i, \sigma_y^2) \) with priors for \( \alpha, \beta \) and \( \sigma_y \).

```r
model {
    for (i in 1:n) {
        y[i] ~ dnorm (y.hat[i], tau.y)
        y.hat[i] <- a + b*x[i]
    }
    a ~ dnorm (0, .0001)
    b ~ dnorm (0, .0001)
    tau.y <- pow(sigma.y, -2)
    sigma.y ~ dunif (0, 100)
}
```

\( \tau_y = 1/\sigma_y^2 \) says \( \tau_y \) is the precision of the \( y_i \).
No pooling and complete pooling

- In complete pooling, we exclude the categorical variable from the model.
- Complete pooling ignores variation between counties.
- In no pooling, we estimate separate models within each group.
- No pooling analysis overfit the data within each county.

Partial pooling: compromise between both extremes.
Partial-pooling estimates

The multilevel estimate for county $j$ is approximately a weighted average of

- the mean of the observations in county $j$, i.e. the “unpooled estimate” $\bar{y}_j$
- and the mean over all counties, i.e. the “completely pooled estimate” $\bar{y}_{all}$

given by

$$\hat{\alpha}_j \approx \frac{n_j \bar{y}_j + \frac{1}{\sigma_b^2} \bar{y}_{all}}{\frac{n_j}{\sigma^2} + \frac{1}{\sigma_b^2}}.$$

The above formula for $\hat{\alpha}_j$ is a weighted average:

- for counties with small $\frac{n_j}{\sigma^2}$, the weighting pulls the estimate towards the overall mean $\bar{y}_{all}$
- while for counties with large $\frac{n_j}{\sigma^2}$, the weighting pulls the estimate towards the group mean $\bar{y}_j$.

We remark that data from group $j$ has an effect on the estimate $\hat{\alpha}_k$ for $k \neq j$ (via its effect on $\mu$ and $\sigma^2_b$).

So there is a “sharing” or “borrowing” of information across groups.
Within-group and between-group regressions

The *within-group regression coefficient* is the regression coefficient within each group, assumed to be the same across the groups.

The *between-group regression coefficient* is defined as the regression coefficient for the regression of the group means of Y on the group means of X.

This distinction is essential to avoid so called *ecological fallacies*.
Ecological fallacies

The ecological fallacy occurs when you make conclusions about individuals based only on analyses of group data.

Example: Score per school

• assume that you measured the score at an exam in various schools and found that a particular school had the highest average score in the district.

• imagine you meet one of the kids from that school and you think to yourself ”she must be very good at this exam !”

• Fallacy! Just because she comes from the school with the highest average doesn’t mean that she automatically obtained a high-scorer in the exam.
Example of model with ecological fallacy

Consider the model

$$y_{ij} = \beta_1 + b_j + \beta_2 x_{ij} + \epsilon_{ij}$$

with one continuous explanatory variable $x$.

This model does not take into account that the within-group coefficient may differ from the between-group coefficient.

- **Within-group coefficient** = expected difference in $y$ between two individuals, in the same group, who differ by 1 unit in $x_{ij}$.
- **Between-group coefficient** = expected difference in group mean $\bar{y}_{.j}$ between two groups which differ by 1 unit in $\bar{x}_{.j}$.
For model:

\[ y_{ij} = \beta_1 + b_j + \beta_2 x_{ij} + \epsilon_{ij} \]

- within group \( j \), the non-random part of the regression is \( y = \beta_1 + \beta_2 x \) and so the \textit{within-group coefficient} is \( \beta_2 \)
- taking the average of \( y_{ij} \) over individuals in group \( j \) gives

\[ \bar{y}.j = \beta_1 + b_j + \beta_2 \bar{x}.j + \bar{\epsilon}.j \]

and the non-random part of this regression is \( \bar{y} = \beta_1 + \beta_2 \bar{x} \), hence the \textit{between-group coefficient} is also \( \beta_2 \).

Remark: the result is the same if we add a random-effect in the slope.
If we add the group mean $\bar{x}_j$ as an explanatory variable, we obtain

$$y_{ij} = \beta_1 + b_j + \beta_2 x_{ij} + \beta_3 \bar{x}_j + \epsilon_{ij}$$

and this allows within- and between-group coefficients to differ:

- within group $j$, the non-random part of the regression is $y = (\beta_1 + \beta_3 \bar{x}_j) + \beta_2 x$ and so the within-group effect is $\beta_2$
- taking the average of $y_{ij}$ over individuals in group $j$ gives

$$\bar{y}_j = \beta_1 + b_j + (\beta_2 + \beta_3) \bar{x}_j + \bar{\epsilon}_j$$

and the non-random part of this regression is $\bar{y} = \beta_1 + (\beta_2 + \beta_3) \bar{x}$, hence the between-group effect is $\beta_2 + \beta_3$.

The difference between the two models can be tested by testing if $\beta_3 = 0$. 
Within- and between-group regressions for IQ

Example from Snijders and Bosker (2012), continued.

We consider

\[ y_{ij} = \beta_1 + b_{j1} + (\beta_2 + b_{j2})IQ_{ij} + \beta_3 \overline{IQ}_j + \epsilon_{ij}. \]

```r
lang2.lme <- lme(langscore ~ 1 + IQ + IQsch.av, 
                 random = ~ 1+IQ | schoolnr, data = nlschools)
```

The variable IQsch.av contains the average IQ score for each child’s school.
```r
> summary(lang2.lme)
....

Random effects:
Formula: ~1 + IQ | schoolnr

<table>
<thead>
<tr>
<th></th>
<th>StdDev</th>
<th>Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2.9964325</td>
<td>(Intr)</td>
</tr>
<tr>
<td>IQ</td>
<td>0.4465464</td>
<td>-0.628</td>
</tr>
<tr>
<td>Residual</td>
<td>6.2996705</td>
<td></td>
</tr>
</tbody>
</table>

Fixed effects: langscore ~ 1 + IQ + IQsch.av

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std.Error</th>
<th>DF</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>41.12752</td>
<td>0.23468112</td>
<td>3546</td>
<td>175.24852</td>
<td>0e+00</td>
</tr>
<tr>
<td>IQ</td>
<td>2.48020</td>
<td>0.06449699</td>
<td>3546</td>
<td>38.45450</td>
<td>0e+00</td>
</tr>
<tr>
<td>IQsch.av</td>
<td>1.03055</td>
<td>0.26334813</td>
<td>209</td>
<td>3.91326</td>
<td>1e-04</td>
</tr>
</tbody>
</table>

The within-group coefficient is 2.48 and the between-group coefficient is 2.48 + 1.03 = 3.51.
• A pupil with a higher IQ score obtains, on average, a higher langscore: on average, 2.48 extra langscore points per additional IQ point.

• In addition, a pupil with a given IQ obtains, on average, a higher langscore if s/he is in a class (i.e. school) with a higher average IQ: if the class average IQ increases by 1 point, then a pupil’s langscore increases by 1.03 points (on average).

So the effect of the classroom average IQ is about half that of the individual’s IQ. The within- and between-group regressions are the same if $\beta_3 = 0$. 
If the within- and between-group coefficients differ, then instead of using $x_{ij}$ and $\bar{x}.j$ it is often convenient to use $x_{ij} - \bar{x}.j$ and $\bar{x}.j$. So then the model is

$$y_{ij} = \gamma_1 + b_j + \gamma_2(x_{ij} - \bar{x}.j) + \gamma_3\bar{x}.j + \epsilon_{ij}$$

$$= \gamma_1 + b_j + \gamma_2\tilde{x}_{ij} + \gamma_3\bar{x}.j + \epsilon_{ij}$$  \hspace{1cm} (1)

where $\tilde{x}_{ij} = x_{ij} - \bar{x}.j$ and where $\gamma_1 = \beta_1$, $\gamma_2 = \beta_2$, $\gamma_3 = \beta_2 + \beta_3$.

The form (1) is convenient as the within-group coefficient $\gamma_2$ and the between-group coefficient $\gamma_3$ (and their standard errors) are given immediately once we have fitted (1).
References

Snijders and Bosker (2012) note that it is common for within- and between-group coefficients to differ, and that their interpretations are different.

See Snijders and Bosker (2012), Sections 3.1, 4.6.

See also:

Available from:
http://www.stat.columbia.edu/~gelman/research/published/

[Rich *states* vote for the Democrats, but rich *people* vote Republican.]
Further hierarchical models

So far, we have focused on hierarchical linear models with two levels.

In real-life situations, very often we need to consider

- more levels
- non-nested models
- non-linear models
A three-level model

Suppose we have data with the structure: students nested within classes, and classes nested within schools. That is, we have 3 levels (students, classes, schools).

This is the structure of dataset science in package DAAG. The response variable like is a measure of a student’s attitude to science on a scale from 1 (dislike) to 12 (like). Explanatory variables available include: sex, a factor with levels f, m; and PrivPub, a factor with levels private school, public school.

```r
## may need to install DAAG first
library(DAAG)

science.lme <- lme(like ~ sex + PrivPub,
                   random = ~ 1 | school/class,
                   data = science, na.action = na.omit)
```
We read “random = ~ 1 | school/class” as “a random intercept for each school, plus a random intercept for each class within each school.”

Non-nested grouping factors

In the above model, each student belongs to exactly one class, and each class belongs to exactly one school – there is a nesting structure.

When grouping factors are not nested, they are said to be “crossed”.

Package lme4 is more suited than nlme to handling non-nested data.

http://cran.r-project.org/web/packages/lme4/index.html
Example of non-nested model

From Gellman et al.

Psychological experiment with two potentially interacting factors:

- psychological experiment on 40 pilots on flight simulators
- under $J = 5$ different treatment conditions
- coming from $K = 8$ different airports

For each pilot $i$, we denote by

- $j[i]$ its treatment condition
- $k[i]$ its associated airport
The response can be fitted to a non-nested model

\[
y_{ij} \sim N \left( \mu + \gamma_{j[i]} + \delta_{k[i]}, \sigma^2 \right) \text{ for } i = 1 \ldots 40
\]

\[
\gamma_{j[i]} \sim N \left( 0, \sigma^2_{\gamma} \right) \text{ for } j = 1 \ldots 5
\]

\[
\delta_{k[i]} \sim N \left( 0, \sigma^2_{\delta} \right) \text{ for } k = 1 \ldots 8
\]

The parameters \( \gamma_{j} \) and \( \delta_{k} \) represent respectively the treatment and airport effect.

When we fit to the data, the estimate residual standard deviations at the individual, treatment and airport levels are \( \sigma = 0.23 \), \( \sigma_{\gamma} = 0.04 \), and \( \sigma_{\delta} = 0.32 \). Thus the variation among the airports is very big but the treatments have almost no impact in the variation.
More complex multilevel models

The models we have considered so far can be generalised in a number of way, in particular within logistic regression or generalised linear models. Other extensions include:

- variance can vary, as parametric functions of input variables and in a multilevel way by allowing different variances for groups
- models with several factors can have many potential interactions, which themselves can be modelled in structured way
- regression models can be set up for multivariate outcomes
- more complicated models are appropriate for spatial data or network structure