Correlated Data: Linear Mixed Models with Random Intercepts
Mixed Effects Models

- This lecture introduces linear mixed effects models.

- Linear mixed models are a type of regression model, which generalise the linear regression model.

- The generalisation allows us to relax the assumption of linear regression, that the errors are independent and therefore uncorrelated.
Outline

▶ Recap of multivariate linear regression.
▶ Motivating example: growth curves: why linear regression isn’t the best model.
▶ General linear mixed model formulation.
▶ The random intercept model.
▶ Prediction using BLUP.
Recap of Linear Regression

- The linear regression model has equation:
  \[ \text{E}Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p \]

- The statistical model for observation \( i \) is:
  \[ Y_i = \text{E}Y_i + \epsilon_i \]
  \[ = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots \beta_p X_{ip} + \epsilon_i \]
  with \( \epsilon_i \sim N(0, \sigma^2) \).

- An assumption of linear regression is that the observed errors are independent and so uncorrelated.

- That is, for any pair of observations \( i_1, i_2 \):
  \[ \text{Cor}(\epsilon_{i_1}, \epsilon_{i_2}) = 0. \]
Motivating Example

- Potthoff and Roy (1964) Biometrika 51:313-326

- Measured distance from the pituitary gland (at the base of the brain) to the pterygomaxillary fissure (point in skull bone) for a sample of 16 males and 11 females.

- Measurements taken from X-rays every 2 years, between the ages of 8 and 14.

- We are interested in modelling growth i.e. in the relationship between distance and age.
Growth Curve Example

- The data are available in the R package **nlme**.
- **nlme** also contains routines that allow us to fit linear mixed effects models.
- This package has been superseded and we will instead use "lme4" for analysis.
- To install the packages (just once):
  ```
  > install.package("nlme")
  > install.package("lme4")
  ```
- To load a package (each time we start R):
  ```
  > library(nlme)
  > library(lme4)
  ```
Growth Curve Example

- The data are in the dataframe Orthodont.

- List the variables in the dataframe:
  
  ```
  > names(Orthodont)
  [1] "distance" "age" "Subject" "Sex"
  ```

- The variables are as follows

<table>
<thead>
<tr>
<th>Variable Description</th>
<th>R name</th>
</tr>
</thead>
<tbody>
<tr>
<td>distance (in mm) from pituitary gland to the pterygomaxillary fissure</td>
<td>distance</td>
</tr>
<tr>
<td>age of subject at time of measurement</td>
<td>age</td>
</tr>
<tr>
<td>Factor variable identifying subject</td>
<td>Subject</td>
</tr>
<tr>
<td>Factor variable identifying sex of subject (Male/Female)</td>
<td>Sex</td>
</tr>
</tbody>
</table>
Growth Curve Example

- Look at the first few rows of the dataframe:

```r
> head(Orthodont)
   distance age Subject Sex
  1     26.0   8     M01  Male
  2     25.0  10     M01  Male
  3     29.0  12     M01  Male
  4     31.0  14     M01  Male
  5     21.5   8     M02  Male
  6     22.5  10     M02  Male
```

- Note: there are multiple observations on each subject.

- The data have a natural grouping structure. There is a group of observations (age and sex measurements) for each study subject.

- There are some variables which vary at the group level but not the observation level e.g. the sex of a subject is fixed across age and distance measurements.
Growth Curve Example

- Because adolescent growth patterns depend strongly on sex to simplify we will restrict analysis to females.

- Create a new dataframe, just containing the observations on female subjects:

  OrthoFem = Orthodont $Sex == "Female", ]
Growth Curve Example: Linear Regression

- Suppose we fit a linear regression, regressing distance on age:

```r
> lm.obj<-lm(distance~age, data=OrthoFem)
> summary(lm.obj)
```

Call:
```
lm(formula = distance ~ age, data = OrthoFem)
```

Residuals:
```
  Min 1Q Median 3Q Max
-4.7091 -1.1784 -0.1068 1.5080 4.8727
```

Coefficients:
```
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            17.3727     1.6378 10.608 1.87e-13 ***
age                     0.4795     0.1459  3.287  0.00205 **
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Each year of age corresponds to about 0.5mm of growth.
Growth Curve Example: Linear Regression

- Plot `distance` against `age`:
  ```r
  > plot(distance~age, data=OrthoFem)
  ```
Growth Curve Example: Linear Regression

- Add a line of best fit from regression of distance on age:
  
  ```R
  > abline(lm.obj$coef)
  ```
Assumptions about Errors

▶ The key assumption of linear regression is that the errors

\[ \epsilon_i = Y_i - \mathbb{E}Y_i \]
\[ = Y_i - \alpha - \beta_1 X_{i1} - \beta_2 X_{i2} \ldots - \beta_p X_{ip} \]

are independent and identically distributed.

▶ Is this assumption realistic for the skull growth data?

▶ One way to examine the errors is to look at the residuals which are estimates of the errors:

\[ \hat{\epsilon}_i = Y_i - \hat{Y}_i \]
\[ = Y_i - \hat{\beta}_1 X_{i1} - \hat{\beta}_2 X_{i2} \ldots - \hat{\beta}_p X_{ip} \]

▶ If the assumptions of linear regression hold errors should not be correlated with the value of the outcome \( Y_i \).
Assumptions about Errors

Plot the residuals $\hat{\epsilon}_i$ against the fitted values $\hat{Y}_i$

> plot(lm.obj$fitted, lm.obj$residuals,
  xlab="Fitted Value", ylab="Residual")
Correlated Errors

- Each coloured line corresponds to a study subject, residuals from the same subject are correlated.
Growth Plots for the 11 Females in the Sample

> plot(OrthoFem)
Correlated Errors

- Overlay the line of best fit on each curve:
Correlated Errors

- Why are the errors from the same subjects correlated?
- Because large 8 year olds are large 14 year olds etc.
- How can we remove this correlation?
- What if we were to allow the intercept to vary across subjects?
Growth Plots for the 11 Females in the Sample

- Fix the slope (growth rate) across subjects but allow different intercepts (red lines):
The Linear Mixed Model

- Correlated errors violate the assumption of the linear model.
- A model that accounts for the error correlation within study subjects should provide a better fit.
- The linear mixed model (LMM) is a generalisation of linear regression, which allows us to model correlated errors.
The Linear Mixed Model

- Similar to multivariate linear regression.

\[ Y_i = \sum_j \beta_j X_{ij} + \sum_k^q u_k Z_{ik} + \epsilon_i \]

- The \( q \) new terms \( u_1, \ldots, u_q \) are called the random effects.

- Like the error term \( \epsilon_i \), the \( u_k \) are random terms.

- The noise terms \( \epsilon_i \), must be independent as in ordinary linear regression.

- The random effects terms \( u_k \) can have any correlation structure we choose.

- The \( Z_{ik} \) are predictor variables which depend on the structure of the experiment. (We will see examples.)
The Random Intercept Model

- A simple hierarchical LMM

- Applies when the data are grouped and the intercept varies between groups (as in the growth example).
The Random Intercept Model

- A simple hierarchical LMM

- Applies when the data are grouped and the intercept varies between groups (as in the growth example).

- We need two indices for the observations, one for each level of the hierarchy.

- $i$ indexes observations while $k$ indexes the group so $k(i)$ is the group of the $i$th observation:

$$Y_i = \alpha + \sum_{j}^{p} \beta_j X_{ij} + u_{k(i)} + \epsilon_i, \quad u_k \sim N(0, \psi^2), \quad \epsilon_i \sim N(0, \sigma^2)$$

This model allows the intercept to shift randomly between the groups.
The Random Intercept Model is a LMM

- The random intercept model:

\[ Y_i = \alpha + \sum_{j}^{p} \beta_j X_{ij} + u_{k(i)} + \epsilon_i, \quad u_k \sim N(0, \psi^2), \quad \epsilon_i \sim N(0, \sigma^2) \]

- The general LMM model:

\[ Y_i = \sum_{j}^{p} \beta_j X_{ij} + \sum_{k}^{q} u_k Z_{ik} + \epsilon_i \]

- We can get the random intercept model as a specialisation of the general LMM model by letting \( q \) be the number of groups and setting \( Z_{ik} = 1 \) if \( i \) is in group \( k \), \( Z_{ik} = 0 \) otherwise.
Fitting The Random Intercept Model to the Growth Data

We use the R function `lmer` which stands for *linear-mixed effects regression*. It is in the `lme4` package.

```r
> lmer(distance~age + (1 | Subject), data=OrthoFem)
```

The syntax for `lmer` is similar to that for `lm`. The first two parameters are the same:
1. a formula giving the fixed effects: `distance~age`
2. a parameter specifying the data-frame: `OrthoFem`

The bit in brackets specifies the random effects:
- `(1 | Subject)` indicates a random intercept
- `| Subject` indicates that there should be a different random effect for each level of the Subject variable.
Fitting The Random Intercept Model to the Growth Data

```r
> rand.int.obj <- lmer(distance ~ age + (1 | Subject), data = OrthoFem)
> rand.int.obj

Linear mixed model fit by REML ['lmerMod']
Formula: distance ~ age + (1 | Subject)
   Data: OrthoFem
REML criterion at convergence: 141.2183
Random effects:
   Groups   Name     Std.Dev.
     Subject (Intercept) 2.068
     Residual            0.780
Number of obs: 44, groups: Subject, 11
Fixed Effects:
(Intercept)     age
       17.3727   0.4795
```
Fitting The Random Intercept Model to the Growth Data

Like \texttt{lm}, \texttt{lmer} reports an estimate of the fixed effect parameters $\beta$: \[
\begin{array}{c}
\text{(Intercept)} \\
17.3727 \\
\end{array} \quad \begin{array}{c}
\text{age} \\
0.4795 \\
\end{array}
\]

It also reports estimates of the variance components parameters $\psi$ and $\sigma$: \[
\begin{array}{c}
\text{Random effects:} \\
\text{Groups} \\
\text{Name} \\
\text{Std.Dev.} \\
\text{Subject} \\
(\text{Intercept}) \\
2.068 \\
\text{Residual} \\
0.780 \\
\end{array}
\]

$\psi = 2.06, \sigma = 0.78$ so that $86\% = \frac{\psi^2}{(\psi^2 + \sigma^2)} \times 100\%$ of the residual variation is explained by the random intercept terms.
Confidence Intervals

- The `confint` command reports confidence intervals:

```r
> confint(rand.int.obj)
Computing profile confidence intervals ...
   2.5 %   97.5 %
.sig01  1.3356536  3.2373519
.sigma  0.6143927  0.9985162
(Intercept)  15.6891967  19.0562579
age     0.3750181  0.5840728
```

- The bottom two lines give the usual 95% confidence intervals for the fixed intercept and random effects.

- The top two lines are approximate confidence intervals for the variance components parameters $\psi$ and $\sigma$. 
Best Linear Unbiased Predictor (BLUP)

- The random effects $u$ are not model parameters but unobserved latent variables.

- Consequently, they cannot be estimated by maximum likelihood, instead a method called BLUP is used.

- BLUP estimates of the random intercepts can be obtained with the `random.effects` command:

```
random.effects(rand.int.obj)
  (Intercept)
  F10  -4.00532866
  F09  -1.47044943
  F06  -1.47044943
  F01  -1.22903236
  F05  -0.02194701
  F07   0.34017860
  F02   0.34017860
  F08   0.70230420
  F03   1.06442981
```
Predicting $Y_i$ for New Observations

As for other regression models, we can use the `predict` function to predict the $Y_i$ for a new observation, for which we know the predictor values of the fixed effects $X_{i1}, \ldots, X_{ip}$.

There are two cases:

1. We want to predict a new value of $Y_i$ for a new observation in a specified group (i.e. subject in the growth example).
2. We want to predict a new value of $Y_i$ for a new observation in a new group, not so far observed. In this case, the intercept for the new group must be predicted, before we predict the $Y_i$. 
Predicting $Y_i$ for a New Observation within a Group

- For an example of the first case, predict the value of the distance for a new measurement on subject F10 in the growth example, taken when she is 16:

```r
predict(rand.int.obj, newdata=data.frame(age=16, Subject="F10"))
```

```
1
21.04013
```
For an example of the second case, predict the value of the distance for a new measurement on a new subject, taken when she is 16:

```r
> predict(rand.int.obj, allow.new.levels=TRUE,
          newdata=data.frame(age=16, Subject="F12"))
1
25.04545
```

- `allow.new.levels=TRUE` indicates that we need to introduce a new group (Subject) which we have called F12 and a new observation within the group.

- Note that the within Subject F10 prediction for distance is lower than the prediction for distance based on a newly predicted subject. This makes sense as F10 is smaller than an average study subject (c.f. prediction curves).
Multi-Level Modelling Example: Split-Plot Experiment

- An agricultural field experiment.
- 3 varieties of oats: *Golden Rain*, *Victory* and *Marvellous*.
- The aim is to investigate the relationship between yield, variety and the dosage of a nitrogen fertiliser.
Multi-Level Modelling Example: Split-Plot Experiment

Experimental Design:

- The field is divided into 6 replicate blocks.
- Each block is divided into 3 plots and a variety of oats is planted in each.
- Each plot is divided into four sub-plots which are fertilised with different doses of nitrogen.
- The observations can be grouped at two levels: by block and by plot within block.
# The Data for Block I

```r
> Oats[1:12,]
     Block Variety nitro yield
   1     I    Victory  0.0  111
   2     I    Victory  0.2  130
   3     I    Victory  0.4  157
   4     I    Victory  0.6  174
   5     I Golden Rain 0.0  117
   6     I  Golden Rain 0.2  114
   7     I Golden Rain 0.4  161
   8     I Golden Rain 0.6  141
   9     I   Marvellous 0.0  105
  10     I   Marvellous 0.2  140
  11     I   Marvellous 0.4  118
  12     I   Marvellous 0.6  156
```
A Multi-Level Random Intercept Model

- The hierarchical experimental design suggests a model with two levels of random effect.

- Let $Y_i$ be the observed yield of oats from the $i$th subplot from the $k(i)$th plot in the $l(k)$th block. We use the model

$$Y_i = \text{nitro}_i \beta_1 + \text{gld}_{k(i)} \beta_2 + \text{vic}_{k(i)} \beta_3 + \text{marv}_{k(i)} \beta_4 + u_{l(k(i))} + v_{k(i)} + \epsilon_i$$

$$u_l \sim N(0, \psi_1^2), \quad v_k \sim N(0, \psi_2^2), \quad \epsilon_i \sim N(0, \sigma^2)$$
A Multi-Level Random Intercept Model

The hierarchical experimental design suggests a model with two levels of random effect.

Let \( Y_i \) be the observed yield of oats from the \( i \)th subplot from the \( k(i) \)th plot in the \( l(k) \)th block. We use the model

\[
Y_i = \text{nitro}_i \beta_1 + \text{gld}_{k(i)} \beta_2 + \text{vic}_{k(i)} \beta_3 + \text{marv}_{k(i)} \beta_4 + u_{l(k(i))} + v_{k(i)} + \epsilon_i
\]

\[
u_l \sim N(0, \psi_1^2), \quad v_k \sim N(0, \psi_2^2), \quad \epsilon_i \sim N(0, \sigma^2)
\]

The fixed effects

- \( \text{nitro}_i \) is the fertiliser dose for the \( i \)th subplot. Question: does fertiliser affect yield i.e. can we reject \( \beta_1 = 0 \)?

- \( \text{gld}_k, \text{vic}_k, \text{marv}_k \) are dummy variables for the variety planted in the \( k \)th plot. Question: does variety affect yield i.e. can we reject \( \beta_2 = \beta_3 = \beta_4 \)?
A Multi-Level Random Intercept Model

The hierarchical experimental design suggests a model with two levels of random effect.

Let $Y_i$ be the observed yield of oats from the $i$th subplot from the $k(i)$th plot in the $l(k)$th block. We use the model

$$Y_i = \text{nitro}_i \beta_1 + \text{gld}_{k(i)} \beta_2 + \text{vic}_{k(i)} \beta_3 + \text{marv}_{k(i)} \beta_4 + u_{l(k(i))} + v_{k(i)} + \epsilon_i$$

The random effects, two levels of random intercept:

- $u_l$ allows for random yield fluctuations between blocks.
- $v_k$ allows for random yield fluctuations between plots (within each block).

These random fluctuations in yields might be due for example to different soil or drainage conditions in different blocks/plots.
Fitting the Multi-Level Model

We can fit the multilevel model in R using the `lme` command

\[ \text{Mod2Levels} <- \text{lmer}(\text{yield} \sim \text{nitro} + \text{Variety} + (1 | \text{Block/Variety}), \text{data=Oats}) \]

Here `Block/Variety` specifies random effects for blocks and random effects for plots within blocks. (We can use `Variety` as a synonym for plot here because within a block each plot corresponds to one and only one variety)
Is the Multi-Level Model Parsimonious?

The multilevel model has random effects $18 + 6 = 24$ random effects. Would it be as effective to use simpler model?

A model with a single set of random effects at the Block level:

```r
> Mod1Level<-lmer(yield~nitro+Variety + (1|Block),
                  data=Oats)
```

An ordinary linear model with no random effects:

```r
> Mod0Levels=lm(yield~nitro+Variety, data=Oats)
```
Is the Multi-Level Model Parsimonious?

Compare using AIC:

```r
> anova(Mod2Levels, Mod1Level, Mod0Levels)
refitting model(s) with ML (instead of REML)
Data: Oats
Models:
Mod0Levels: yield ~ nitro + Variety
Mod1Level: yield ~ nitro + Variety + (1 | Block)
Mod2Levels: yield ~ nitro + Variety + (1 | Block/Variety)

Df   AIC   BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
Mod0Levels 5 650.23 661.61  -320.11  640.23
Mod1Level 6 620.80 634.46  -304.40  608.80  31.428 1 2.069e-08 ***
Mod2Levels 7 615.11 631.04  -300.55  601.11  7.690 1 0.005553 **
```

The two level model has the lowest AIC.
Comparing Models with Different Fixed Effects

> Mod2Levels_no_nitro<-lmer(yield~Variety + (1|Block/Variety), data=Oats)

> Mod2Levels_no_variety<-lmer(yield~nitro + (1|Block/Variety), data=Oats)

> Mod2Levels_no_nitro_no_variety<-lmer(yield~ + (1|Block/Variety), data=Oats)
Comparing Models with Different Fixed Effects

> anova(Mod2Levels_no_nitro_no_variety, Mod2Levels_no_variety, Mod2Levels_no_nitro, Mod2Levels)
refitting model(s) with ML (instead of REML)
Data: Oats
Models:
Mod2Levels_no_nitro_no_variety: yield ~ +(1 | Block/Variety)
Mod2Levels_no_variety: yield ~ nitro + (1 | Block/Variety)
Mod2Levels_no_nitro: yield ~ Variety + (1 | Block/Variety)
Mod2Levels: yield ~ nitro + Variety + (1 | Block/Variety)

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>deviance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mod2Levels_no_nitro_no_variety</td>
<td>4</td>
<td>675.48</td>
<td>684.59</td>
<td>-333.74</td>
<td>667.48</td>
</tr>
<tr>
<td>Mod2Levels_no_variety</td>
<td>5</td>
<td>614.23</td>
<td>625.61</td>
<td>-302.11</td>
<td>604.23</td>
</tr>
<tr>
<td>Mod2Levels_no_nitro</td>
<td>6</td>
<td>676.37</td>
<td>690.03</td>
<td>-332.19</td>
<td>664.37</td>
</tr>
<tr>
<td>Mod2Levels</td>
<td>7</td>
<td>615.11</td>
<td>631.04</td>
<td>-300.55</td>
<td>601.11</td>
</tr>
</tbody>
</table>

- There is good evidence that nitrogen fertiliser has an effect on yield
- There does not seem to be enough evidence to imply that any one variety produces different yields from the other two.
Summary

- We’ve had an initial look at linear mixed effects models (LMMs)
- These allow us to model data with correlated errors
- We’ve looked at the specific case of random intercept models
- These are models were the error correlation is induced by a grouping of observations
- Within each grouping there is a slightly different linear regression intercept