Class 5: Linear mixed models

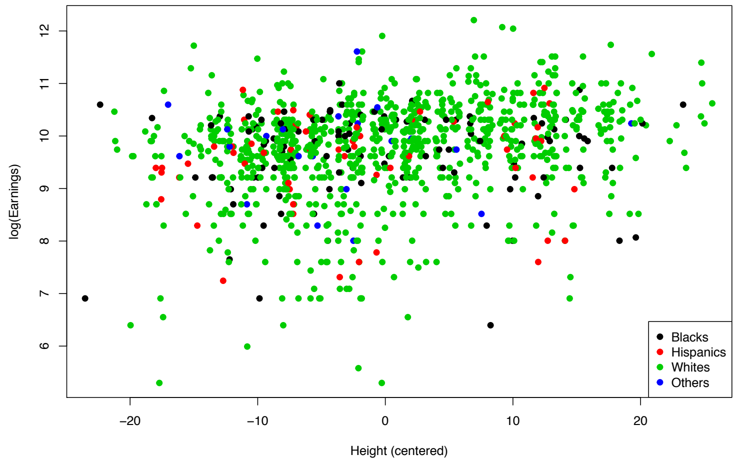
Andrew Parnell
andrew.parnell@mu.ie

Maynooth University
National University of Ireland Maynooth

Learning outcomes

- Understand more complex linear mixed models
- Fit some linear mixed models of different types
- Understand the `lme4` (and `rstanarm`) formula construction
- Know how to do basic model comparison

Reminder: earnings data



First `lme4` model

```
mm_1 = lmer(y ~ x_centered + (1 | eth), data = dat)
summary(mm_1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ x_centered + (1 | eth)
##    Data: dat
##
## REML criterion at convergence: 2810.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.9009 -0.4088  0.1808  0.6309  2.5630
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  eth      (Intercept) 0.001828 0.04275
##  Residual             0.821243 0.90622
## Number of obs: 1059, groups:  eth, 4
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 9.720272   0.042502 228.703
## x_centered  0.022464   0.002868   7.833
##
## Correlation of Fixed Effects:
##            (Intr)
## x_centered 0.023
```

Adding in variable slopes

- This model forces all the fitted lines to be parallel
- i.e. each different ethnic group has the same height/earnings relationship, but shifted up or down
- We can also fit a model with varying slopes

Varying slopes model

```
mm_2 = lmer(y ~ (x_centered | eth), data = dat,
            control = lmerControl(optimizer = "Nelder_Mead"))
summary(mm_2)
```

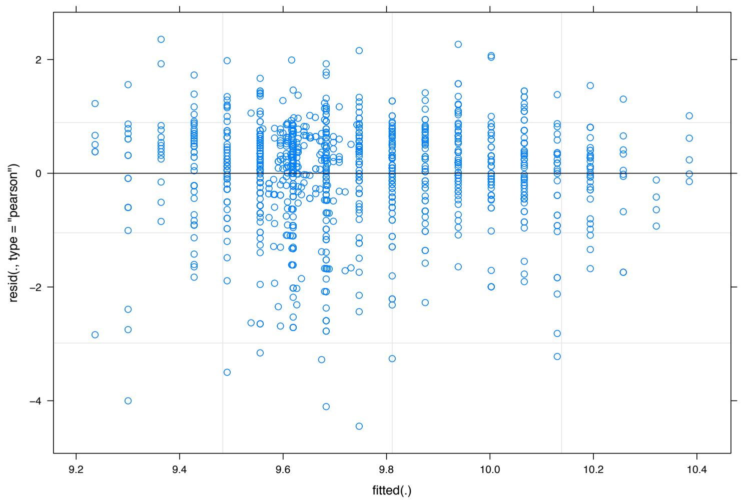
```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ (x_centered | eth)
##    Data: dat
## Control: lmerControl(optimizer = "Nelder_Mead")
##
## REML criterion at convergence: 2800.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.9250 -0.4258  0.1900  0.6467  2.6079
##
## Random effects:
##  Groups   Name        Variance  Std.Dev. Corr
##  eth      (Intercept) 0.0069542 0.08339
##           x_centered  0.0002438 0.01561  1.00
##  Residual             0.8159788 0.90332
## Number of obs: 1059, groups:  eth, 4
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  9.61727    0.03149   305.4
```


A note about 'optimizers'

- `lme4` fits the models via REstricted Maximum Likelihood (REML)
- This parts of the model first (the fixed effects part) and then the random effects second
- `lme4` comes with a variety of methods for maximising the likelihood. The default is 'bobyqa' which seems to fail occasionally. Changing it to one of the other methods often solves the problem

Interpreting output 1

```
plot(mm_2)
```



Interpreting output 2

```
coef(mm_2)
```

```
## $eth
##      x_centered (Intercept)
## 1  0.0044750410    9.641176
## 2  0.0035239801    9.636096
## 3  0.0251281010    9.751489
## 4 -0.0002730401    9.615815
##
## attr(,"class")
## [1] "coef.mer"
```

- Varying intercepts and quite strongly varying slopes

Going back to the maths of the different models

- I always find it helpful to write out the mathematical details of the models I am fitting
- The first model (with varying intercepts) had:

$$y_{ij} = \alpha_j + \beta x_{ij} + \epsilon_{ij}$$

  with $\alpha_j \sim N(\mu_\alpha, \sigma^2_\alpha)$ and $\epsilon_{ij} \sim N(0, \sigma^2)$.
- The second model (with varying intercepts and slopes) has:

$$y_{ij} = \alpha_j + \beta_j x_{ij} + \epsilon_{ij}$$

  with $\alpha_j \sim N(\mu_\alpha, \sigma^2_\alpha)$, $\beta_j \sim N(\mu_\beta, \sigma^2_\beta)$ and $\epsilon_{ij} \sim N(0, \sigma^2)$.
- The key job is to always go back and check that you can match the parameters to the estimates in the output

A third model

- Could also have varying slopes and identical intercepts

```
mm_3 = lmer(y ~ 1 + (x_centered - 1 | eth), data = dat,
            control = lmerControl(optimizer = "Nelder_Mead"))
summary(mm_3)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ 1 + (x_centered - 1 | eth)
##    Data: dat
## Control: lmerControl(optimizer = "Nelder_Mead")
##
## REML criterion at convergence: 2803.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.8971 -0.4310  0.1722  0.6330  2.6255
##
## Random effects:
##  Groups   Name       Variance  Std.Dev.
##  eth      x_centered 0.0002601 0.01613
##  Residual            0.8180663 0.90447
## Number of obs: 1059, groups:  eth, 4
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  9.73190    0.02786   349.4
```

- Can you write out the maths for this model and pick out the parameters?

The dirty secret behind mixed effects models

- Whilst you seem to have a large number of observations (1059 for the earnings example) you only really have 4 observations on the random effects from each of the 4 groups
- Thus the estimates of e.g. the random effect standard deviation is likely to be highly noisy and hard to estimate
- The confidence interval is likely to be big and any assumptions about the distributions of the random effects are likely to be hard to test

What the formulae mean in `lme4`

Formula	Meaning
‘y ∼ x‘	‘y‘ is the response variable, ‘x‘ the single fixed effect with an intercept included too
‘y ∼ x - 1‘	As above but without an intercept term (i.e. $\hat{y} = 0$ when $x = 0$)
‘y ∼ x + (1 | z)‘	As above but with a varying random effect intercept terms grouped by ‘z‘
‘y ∼ (x | z)‘	As above but with varying intercepts and slopes
‘y ∼ 1 + (x - 1 | z)‘	Random effects for slopes but not intercepts

- There are more complicated formulae which we will come on to later

Which model fits best?

- There are lots of ways to compare models, and we will talk more about this later in the course
- `lmer` implements a simple `anova` method for comparing between two models

```
anova(mm_1, mm_2, mm_3)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: dat
## Models:
## mm_3: y ~ 1 + (x_centered - 1 | eth)
## mm_1: y ~ x_centered + (1 | eth)
## mm_2: y ~ (x_centered | eth)
##      Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## mm_3  3 2804.0 2818.9 -1399.0   2798.0
## mm_1  4 2803.5 2823.4 -1397.8   2795.5 2.4552      1     0.1171
## mm_2  5 2805.5 2830.3 -1397.7   2795.5 0.0358      1     0.8500
```


A quick primer on model comparison

- AIC and BIC are commonly used as they balance the fit of the model (measured by the *deviance*) with the complexity of the model
- Lower values of AIC and BIC tend to indicate a better model
- These are methods for ranking models but not good for telling you whether your best model fits well!
- The deviance can also be used to do a chi-squared test; the *p*-value is in the last column

A more complicated model

- The earnings data set also has age group in it (1 = 18-34, 2 = 35-49, and 3 = 50-64).
- Some model ideas:
- Common slope but varying intercepts by both age group and ethnicity
- Varying slopes by age group and varying intercept by ethnicity
- Varying slopes by ethnicity and varying intercepts by age group
- Varying slopes and intercepts by both
- Key task: without looking ahead see if you can write out (a) the maths and (b) the `lme4` formula code for each model

Fitting one of the more complicated models

```r
mm_4 = lmer(y ~ x_centered + (1 | eth) + (1 | age), data = dat,
            control = lmerControl(optimizer = "Nelder_Mead"))
summary(mm_4)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ x_centered + (1 | eth) + (1 | age)
##    Data: dat
## Control: lmerControl(optimizer = "Nelder_Mead")
##
## REML criterion at convergence: 2752.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.7863 -0.4536  0.1752  0.6494  2.8637
##
## Random effects:
##  Groups   Name        Variance  Std.Dev.
##  eth      (Intercept) 0.0007454 0.0273
##  age      (Intercept) 0.0613025 0.2476
##  Residual             0.7730075 0.8792
## Number of obs: 1059, groups:  eth, 4; age, 3
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 9.771905   0.147507  66.247
## x_centered  0.023852   0.002788   8.557
##
## Correlation of Fixed Effects:
##            (Intr)
## x_centered 0.008
```


Comparing this model

```r
anova(mm_1, mm_3, mm_4)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: dat
## Models:
## mm_3: y ~ 1 + (x_centered - 1 | eth)
## mm_1: y ~ x_centered + (1 | eth)
## mm_4: y ~ x_centered + (1 | eth) + (1 | age)
##      Df    AIC    BIC  logLik deviance   Chisq Chi Df Pr(>Chi
## mm_3  3 2804.0 2818.9 -1399.0   2798.0
## mm_1  4 2803.5 2823.4 -1397.8   2795.5  2.4552      1     0.1
## mm_4  5 2750.6 2775.5 -1370.3   2740.6 54.8770      1  1.283e
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```


Interaction effects

- You might have fitted models previously using `lm` with interaction effects

```r
lm(y ~ x*z)
```

- When you create models with varying slopes and intercepts you are really creating an interaction model between the continuous covariate (height in our example) and the categorical covariate (ethnicity or age group)
- Remember always that the mixed effects model has the extra constraint that the different groups (or interaction effects) are tied together

Mathematics for complex mixed effects models

- When we add more variables into the model we need more subscripts

$$y_{ijk} = \alpha_j + \gamma_k + \delta x_{ijk} + \epsilon_{ijk}$$

- Now y_{ijk} is the log earnings value for observation i in ethnic group j and age group k.
- This model has common slope but random intercepts for age group and ethnic group
- We have the extra constraints $\alpha_j \sim N(\mu_\alpha, \sigma^2_\alpha)$, $\gamma_k \sim N(\mu_\gamma, \sigma^2_\gamma)$, $\epsilon_{ijk} \sim N(0, \sigma^2)$

Summary

- We've seen some more complicated models for the earnings data set
- We know how to fit models with multiple mixed effects
- We've used `anova` to compare between models
- We can see how the mathematics and the formula in `lme4` match together