Class 4: Introduction to mixed models

Andrew Parnell
andrew.parnell@mu.ie

Maynooth University
National University of Ireland Maynooth

Learning outcomes

- Know the difference between a simple linear regression and a simple mixed model
- Be able to identify and understand the key features of a mixed model
- Know how to fit a simple mixed model in `lme4`
- Be able to interpret the output of a simple mixed model

What is a mixed effects model?

- You are probably used to seeing *fixed effects* models:

$$y_i = \alpha + \beta x_i + \epsilon_i$$

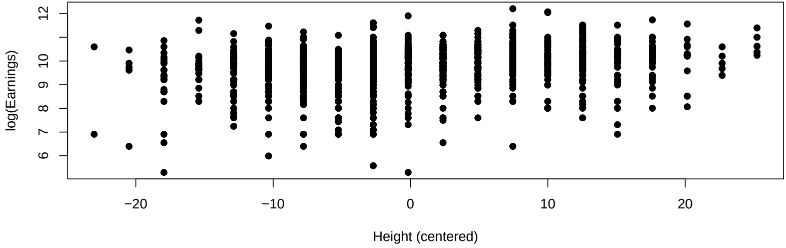
- Here α and β are fixed effects
- That means they do not vary by group (or observation)
- When we add in terms that vary by group or observation and give these a specified probability distribution then we have a *mixed effects* model. You need a categorical covariate to do that.
- (In fact ϵ_i can be considered a *random effect* because it varies by observation and has a constrained distribution $\epsilon_i \sim N(0, \sigma^2)$)

Example data set

- Let's think again about the earnings data where we want to estimate log(earnings) from people's height in cm using a linear regression model where height is mean centered, i.e.

$$\text{log}(\text{earnings}_i) = \alpha + \beta \times (\text{height}_i - \text{mean}(\text{height})) + \epsilon_i$$

```r
dat = read.csv('../data/earnings.csv')
with(dat, plot(x_centered, y, xlab = 'Height (centered)',
     ylab = 'log(Earnings)', pch = 19))
```



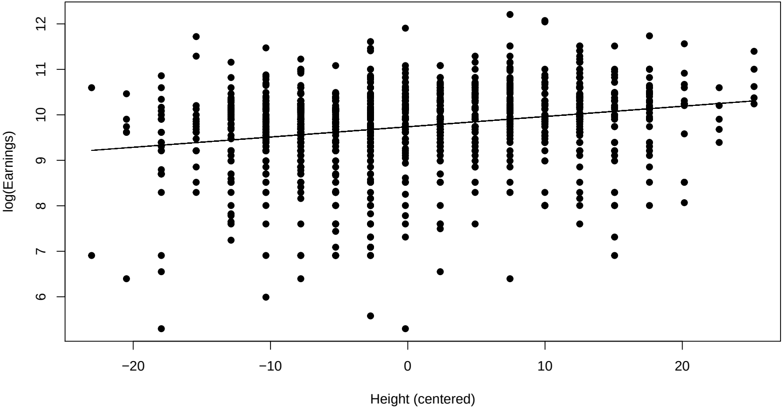
Model fit

```r
summary(lm(y ~ x_centered, data = dat))
```

```
##
## Call:
## lm(formula = y ~ x_centered, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4351 -0.3705  0.1615  0.5761  2.3302
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.737358   0.027858 349.540  < 2e-16 ***
## x_centered  0.022555   0.002866   7.869 8.84e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 0.9066 on 1057 degrees of freedom
## Multiple R-squared:  0.05533,    Adjusted R-squared:  0.05444
## F-statistic: 61.91 on 1 and 1057 DF,  p-value: 8.836e-15
```

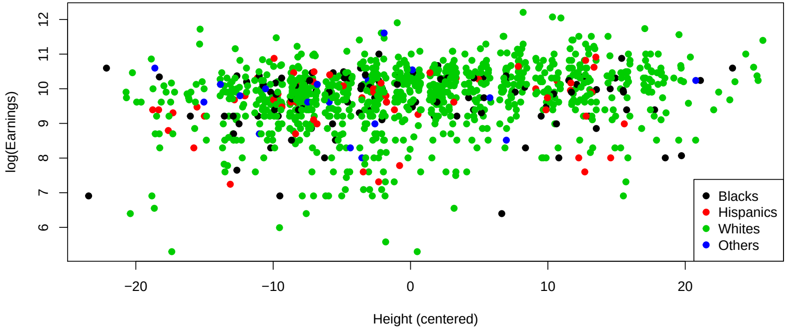
Plot with fitted line

```
with(dat, plot(x_centered, y, xlab = 'Height (centered)',
     ylab = 'log(Earnings)', pch = 19))
lines(dat$x_centered, lm(y ~ x_centered,
                         data = dat)$fitted.values)
```



Using slightly more information

```
eth_names = c('Blacks','Hispanics','Whites','Others')
with(dat, plot(x_centered, y, xlab = 'Height (centered)',
     ylab = 'log(Earnings)', type = 'n'))
with(dat, points(jitter(x_centered, 2), y, col = eth,
                 pch = 19))
legend('bottomright', eth_names, col = 1:4, pch = 19)
```



A new model

Suppose we wanted to fit a simple model where there was a different (parallel) fitted line for each ethnic group:

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

- Note the change of notation. We now write y_{ij} as the ith observation in group (ethnicity) j
- There are 4 ethnicity groups so $j = 1, \ldots, 4$ but different numbers of observations in each group

```
table(dat$eth)
```

```
##
##   1   2   3   4
## 104  61 873  21
```

How could we fit this new model?

I can think of three obvious ways:

1. Divide the data up into 4 groups and fit each individually
2. Fit a linear regression for all the data and include ethnicity as a fixed categorical effect
3. Fit a mixed effects regression model with ethnicity as a random effect

What are the advantages and disadvantages of each?

Fit using lm

```r
summary(lm(y ~ x_centered + as.factor(eth), data = dat))
```

```
##
## Call:
## lm(formula = y ~ x_centered + as.factor(eth), data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4533 -0.3711  0.1599  0.5695  2.3086
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.651865   0.088904 108.565  < 2e-16 ***
## x_centered       0.022328   0.002878   7.759 2.02e-14 ***
## as.factor(eth)2 -0.061071   0.146287  -0.417    0.676
## as.factor(eth)3  0.103612   0.094069   1.101    0.271
## as.factor(eth)4  0.181377   0.217105   0.835    0.404
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
```

A fit using lme

Alternatively we can use lme4 to fit a mixed effects model here:

```
library(lme4)
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
```

Look at the effects for each group

```r
coef(mm_1)
```

```
## $eth
##   (Intercept) x_centered
## 1    9.707429 0.02246425
## 2    9.704834 0.02246425
## 3    9.743485 0.02246425
## 4    9.725341 0.02246425
##
## attr(,"class")
## [1] "coef.mer"
```

- Compare (after a bit of calculation) with the fixed effects model and they should be much more similar to each other

Why are these two models different?

- The `lmer` model has an extra constraint that:

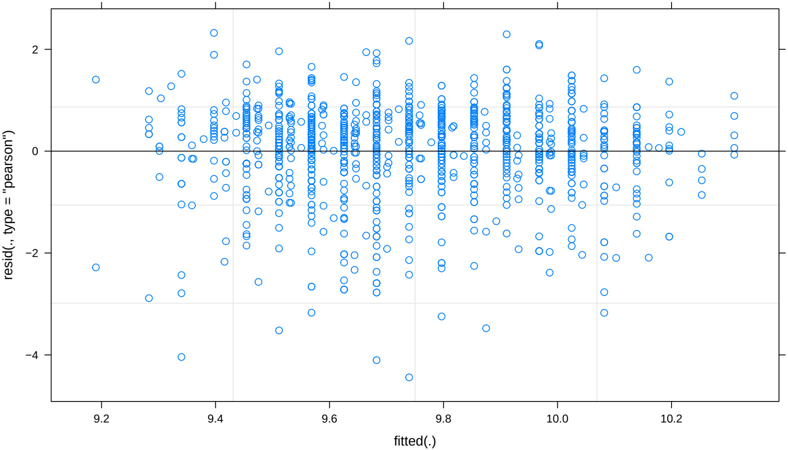
$$\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2)$$

- The constraint forces the intercepts to be tied together
- This has two advantages:
  - We get to borrow strength between groups and reduce the effect of tiny (and noisy) sample sizes (look at the standard errors of the intercepts on the fixed effects version)
  - We can remove the effect of ethnicity from the overall model because we now have an extra estimate of the variability associated with it, via σ_α

Extra plots

The `lme4` package also creates other plots for us:

```r
plot(mm_1)
```



Further output

- Confidence intervals

```
confint(mm_1, level = 0.5)
```

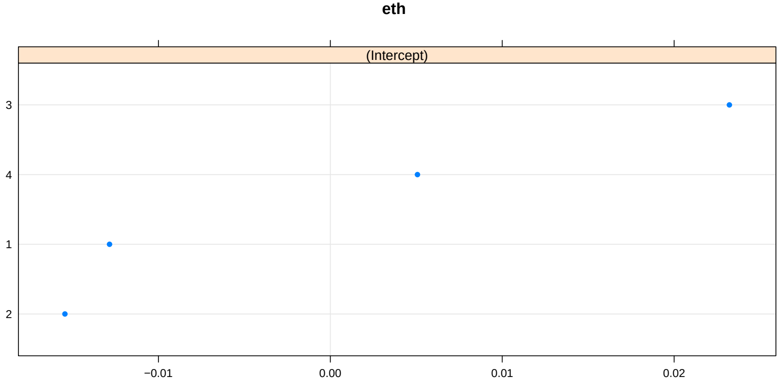
```
## Computing profile confidence intervals ...
```

```
##                   25 %       75 %
## .sig01       0.00000000 0.03164932
## .sigma       0.89258219 0.91913323
## (Intercept) 9.71858380 9.75613175
## x_centered  0.02062318 0.02448677
```

Plot the random effects

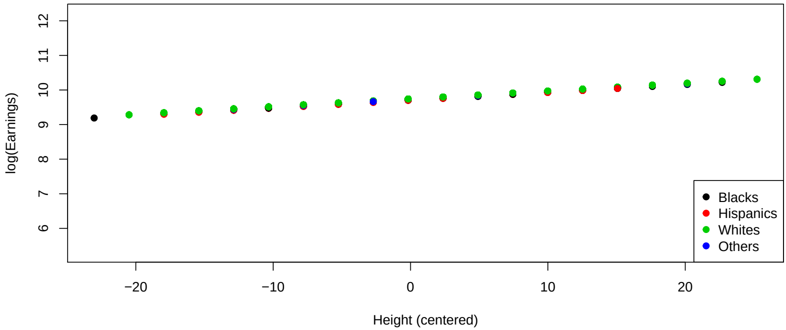
```r
library(lattice)
dotplot(ranef(mm_1))
```

```
## $eth
```



Creating predictions - fiddly

```
fitted_values = predict(mm_1)
with(dat, plot(x_centered, y, xlab = 'Height (centered)',
     ylab = 'log(Earnings)', type = 'n'))
legend('bottomright', eth_names, col = 1:4, pch = 19)
with(dat, points(x_centered, fitted_values, col = eth, pch
```



Summary

- We now know the difference between a *fixed effects* model and a *mixed effects* model
- We have seen some of the key advantages of moving from fitting separately to modelling it jointly
- We have fitted a model in lme4
- We have interpreted some of the lme4 output