# Lecture 1 - fixed vs random effects models

*Prof. Richard Wilkinson*

Last semester you studied what are known as **fixed effects** linear models. In this module, we're going to look at **random effects** models, and **mixed effects** models, which include fixed and random effects. There is no universally accepted definition of what the difference is between a fixed and random effect - see http://andrewgelman.com/2005/01/25/why_i_dont_use/. I will mainly model effects as fixed if they are of interest in themselves, or as random, if interest lies instead in the underlying population. However, note that the choice of appropriate model is (somewhat) a matter of personal choice and judgement. There are situations where some statistians may prefer to model an effect as fixed, but where others would argue it should be random. They key thing is to not get too caught up in the terminology. Jargon should never be used as a substitute for a mathematical understanding of the models!

Let's begin by considering a simple dataset on crop yields. An agricultural company has a new GM wheat variant that it wishes to test. It send the seeds to 6 different farmers, who grew the crop for 5 consecutive years. Each year the crop yield is measured at each farm.

```
str(Cropdata)
```

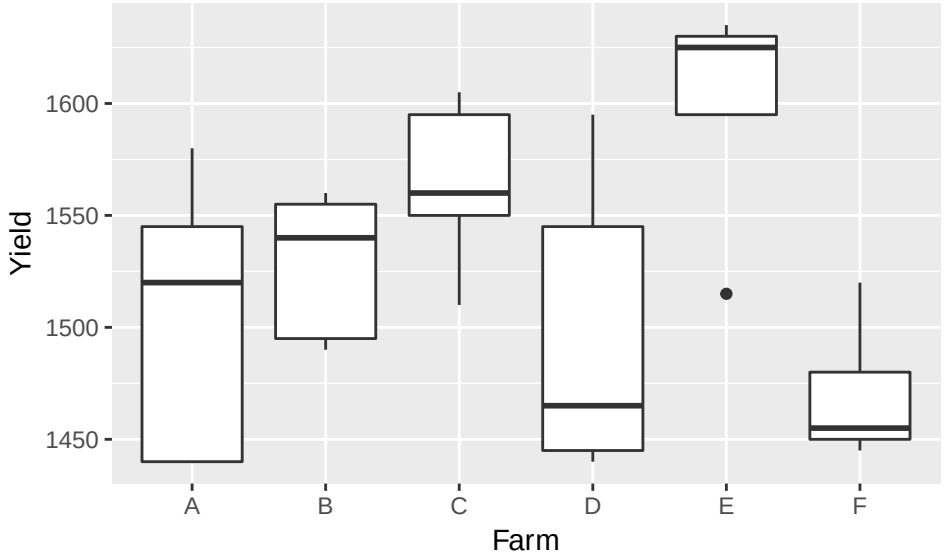
```
## 'data.frame':    30 obs. of  2 variables:
##  $ Farm : Factor w/ 6 levels "A","B","C","D",..: 1 1 1 1 1 2 2 2 2 2 ...
##  $ Yield: num  1545 1440 1440 1520 1580 ...
```

```
head(Cropdata)
```

```
##   Farm Yield
## 1    A  1545
## 2    A  1440
## 3    A  1440
## 4    A  1520
## 5    A  1580
## 6    B  1540
```

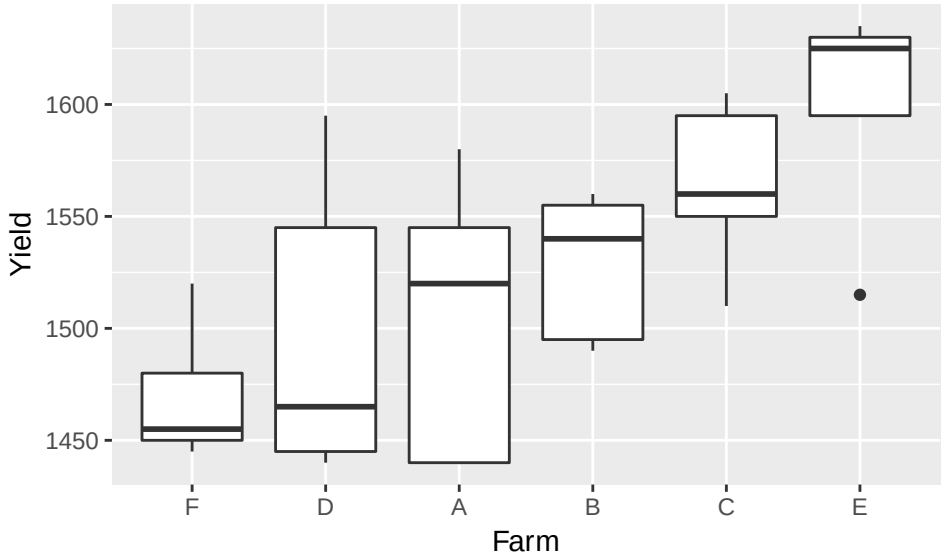
As **always**, we begin by plotting the data. This helps us to understand the structure of the data, and may suggest sensible models. I will use the ggplot2 package here, as it produces clean and elegant plots, but there are many other options for producing similar plots in R.

```
library(ggplot2)
qplot(Farm, Yield, geom='boxplot', data = Cropdata)
```




This would probably look better if we reordered the farms in order of increasing yield.

```
qplot(reorder(Farm, Yield), Yield, geom='boxplot', data = Cropdata, xlab = 'Farm')
```




**Q: What model would you fit?**

Let's begin by fitting a **fixed effects** model of the type you are familiar with. Because we only have one covariate, which is a factor, this type of model is often called a *one-way ANOVA* model.

$$Yield_{ij} = \mu_i + \epsilon_{ij} \text{ for Farm i}$$

```
(fit1 <- lm(Yield~Farm-1, data=Cropdata))
```

```
##
## Call:
## lm(formula = Yield ~ Farm - 1, data = Cropdata)
##
## Coefficients:
## FarmA  FarmB  FarmC  FarmD  FarmE  FarmF
##  1505   1528   1564   1498   1600   1470
```

```
#model.matrix(fit1) # useful for checking what model we have fit.
```

Note that we have had to specify the -1 in the formula to avoid the addition of an intercept. The command

```
lm(Yield~Farm, data=Cropdata)
```

would have fit the model

$$Yield_{ij} = \begin{cases} \alpha + \epsilon_{ij} \text{ for Farm A} \\ \alpha + \mu_i + \epsilon_{ij} \text{ for Farm i} \neq A \end{cases}$$

instead. Alternatively, we could have used the command

```
(fit2 <- lm(Yield~Farm, data=Cropdata, contrasts=list(Farm=contr.sum)))
```

```
##
## Call:
## lm(formula = Yield ~ Farm, data = Cropdata, contrasts = list(Farm = contr.sum))
##
## Coefficients:
## (Intercept)        Farm1        Farm2        Farm3        Farm4
##      1527.5        -22.5          0.5         36.5        -29.5
##       Farm5
##        72.5
```

to fit the model

$$Yield_{ij} = \alpha + \beta_i + \epsilon_{ij} \text{ for Farm i}$$
$$\sum \beta_i = 0, \qquad \epsilon_{ij} \sim N(0, \sigma^2).$$

- Is this fixed effects model useful in this situation? What might we wish to use the model for?
- How would we predict crop yield at some other farm?

## Random effects model

The model above calculates the within farm variance - this is the residual variance (estimated to be $49.5^2$ - type `summary(fit2)$sigma` into R).

- What is the between farm variance?
- What will the mean of the yield be at some other farm not included in the study?

This is a case where random effects models are useful. We don't really care about the different farm means - those farms will never occur again. We are interested in farms not in the sample - in other words, we want to know the distribution of possible yields at different farms.

This is where **random effects** models are useful. We want to model the distribution of the yields at different farms, not compare the yield at the 6 farms in the study.

$$Yield_{ij} = \alpha + b_i + \epsilon_{ij} \text{ for Farm i}$$
$$\text{where } b_i \sim N(0, \sigma_b^2) \text{ and } \epsilon_{ij} \sim N(0, \sigma^2)$$

```
library(lme4)
(fm02 <- lmer(Yield ~ 1+(1|Farm), Cropdata))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Yield ~ 1 + (1 | Farm)
##    Data: Cropdata
## REML criterion at convergence: 319.6543
## Random effects:
##  Groups   Name        Std.Dev.
##  Farm     (Intercept) 42.00
##  Residual             49.51
## Number of obs: 30, groups:  Farm, 6
## Fixed Effects:
## (Intercept)
##        1528
```

Reading from the output above, we can see that the fitted model is

$$Yield_{ij} = 1528 + b_i + \epsilon_{ij}$$

where

$$b_i \sim N(0, 42^2) \text{ and } \epsilon_{ij} = N(0, 49.5^2).$$

- The within-farm variance, is $49.5^2$ as in the fixed effects model.
- But now we have a model for the between-farm variance, estimated here to be $42.0^2$.
- We can use the random effects model to predict the yield at new farms. It will be

$$Yield \sim N(1528, 49.5^2 + 42.0^2)$$

The `lmer` function is used to fit random and mixed effects models. The fixed effects are specified in exactly the same way as they are when using the `lm` command; The random effects are specified within brackets. The term on the left of the `|` gives the model formula, and the term on the right of the `|` describes how the data should be grouped (in this case into farms).

## Comparing fixed and random effects

Benjamin Bolker, in the book Ecological Statistics, makes the following points about random effects models:

"*Frequentists and Bayesians define random effects somewhat differently, which affects the way they use them. Frequentists define random effects as categorical variables whose levels are chosen at random from a larger population, e.g., species chosen at random from a list of endemic species. Bayesians define random effects as sets of variables whose parameters are all drawn from the same distribution.*"

"*Random effects can also be described as predictor variables where you are interested in making inferences about the distribution of values (i.e., the variance among the values of the response at different levels) rather than in testing the differences of values between particular levels.*"

Lets compare the fitted fixed and random effects models.

```
summary(fm02)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Yield ~ 1 + (1 | Farm)
##    Data: Cropdata
##
## REML criterion at convergence: 319.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.4117 -0.7634  0.1418  0.7792  1.8296
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Farm     (Intercept) 1764     42.00
##  Residual             2451     49.51
## Number of obs: 30, groups:  Farm, 6
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  1527.50      19.38    78.8
```

```
summary(fit2)
```

```
##
## Call:
## lm(formula = Yield ~ Farm, data = Cropdata, contrasts = list(Farm = contr.sum))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -85.00 -33.00   3.00  31.75  97.00
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1527.500      9.039 168.985  < 2e-16 ***
## Farm1        -22.500     20.212  -1.113  0.27666
## Farm2          0.500     20.212   0.025  0.98047
## Farm3         36.500     20.212   1.806  0.08351 .
## Farm4        -29.500     20.212  -1.459  0.15739
## Farm5         72.500     20.212   3.587  0.00149 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 49.51 on 24 degrees of freedom
## Multiple R-squared:  0.4893, Adjusted R-squared:  0.3829
## F-statistic: 4.598 on 5 and 24 DF,  p-value: 0.004398
```

Note that the standard error for the estimate of $\alpha$ in the fixed effects model is 9.04, whereas it is 19.38 in the random effects model. Why is it larger in the random effect model?

The predicted random effects, the $\hat{b}_i$, can be found with the `ranef` command:

```
ranef(fm02)
```

```
## $Farm
##   (Intercept)
## A -17.6068514
## B   0.3912634
## C  28.5622255
## D -23.0845384
## E  56.7331877
## F -44.9952868
```

These are similar to the estimated fixed effects, $\hat{\beta}_i$, but they are not the same, as there is some 'shrinkage' from the least squares estimates towards zero. This is a consequence of modelling the random effects as random variables with expectation zero.

# Mixed effect models

Lets now consider the sleepstudy dataset from the lme4 package. This is from a report on a study of the effects of sleep deprivation on reaction time for a number of subjects chosen from a population of long-distance truck drivers. These subjects were divided into groups that were allowed only a limited amount of sleep each night. We consider the group of 18 subjects who were restricted to three hours of sleep per night for the first ten days of the trial. Each subject's reaction time was measured several times on each day of the trial.

```
library(lme4)
str(sleepstudy)
```

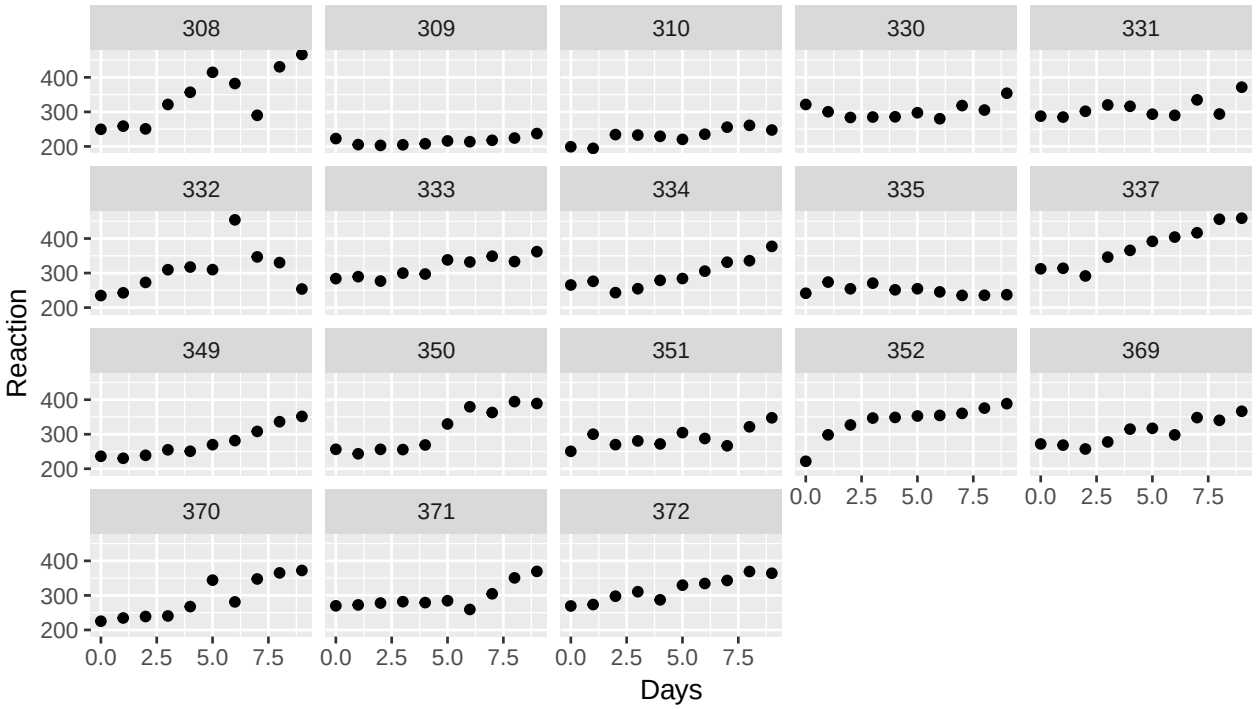
```
## 'data.frame':    180 obs. of  3 variables:
##  $ Reaction: num  250 259 251 321 357 ...
##  $ Days    : num  0 1 2 3 4 5 6 7 8 9 ...
##  $ Subject : Factor w/ 18 levels "308","309","310",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
head(sleepstudy)
```

```
##   Reaction Days Subject
## 1 249.5600    0     308
## 2 258.7047    1     308
## 3 250.8006    2     308
## 4 321.4398    3     308
## 5 356.8519    4     308
## 6 414.6901    5     308
```

As always, we should start by plotting the data.

```
library(ggplot2)
qplot(Days, Reaction, facets=~Subject, data = sleepstudy)
```




We can immediately see that there is a trend for the reaction time to slow down as the trial progresses. We can also see that each individual is affected differently, both in the trend and the initial reaction time

(intercept).

Note that this is a type of longitudinal data, in that the data are repeated measurements on the same subject taken over time.

**What model would you fit?**

- What fixed effects?
- What random effects?
- What combination of fixed and random effects?