Mixed-effects models

An introduction by Christoph Scherber

Up to now, we have been dealing with linear models of the form

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where ß0 and ß1 are parameters of fixed value.

Example:

Let us assume that we are measuring the yield of a crop plant on 5 different plots at 4 different observation times.

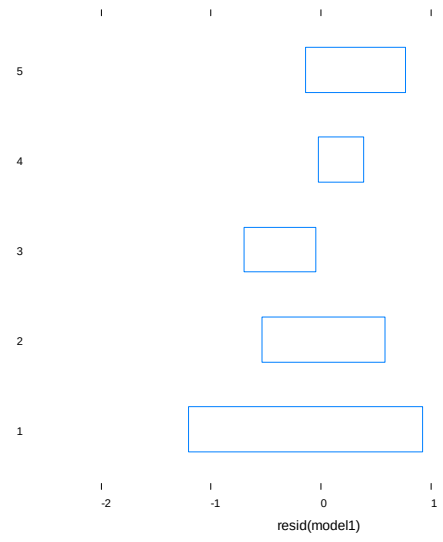
```
yield=rnorm(20,150)

plot=gl(5,4)
```

Let us start off with a wrong model, ignoring the grouping of our data points, and assuming that all 20 plants harvested were independent random samples:

```
Call:
lm(formula = yield ~ 1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.42295 -0.22550  0.06235  0.57603  1.63489

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 149.8674     0.1881   796.8   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8412 on
19 degrees of freedom
```

We conclude that $\widehat{ß}$ = 149. 8674 and $\hat{\sigma}$=0.8412.

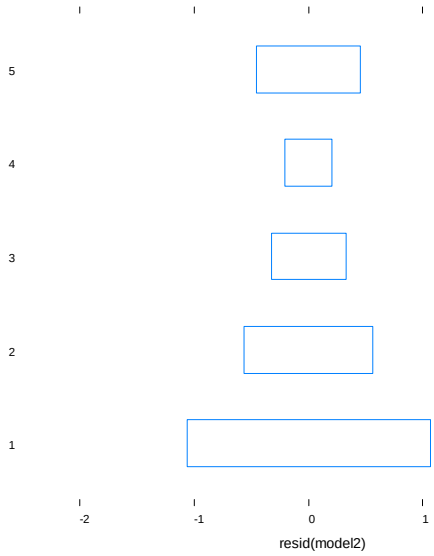
If we inspect the residuals of this model, separately for each plot, we see that there is high variability between plots (right).



We can improve our initial model by formulating a fixed-effects model with a different mean estimated for every plot:

```
model2=lm(yield~plot-1)
summary(model2)

Call:
lm(formula = yield ~ plot - 1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.27885 -0.33863  0.08602  0.46127  1.77899

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
plot1 149.7233     0.4524   331.0   <2e-16 ***
plot2 149.8896     0.4524   331.4   <2e-16 ***
plot3 149.4947     0.4524   330.5   <2e-16 ***
plot4 150.0474     0.4524   331.7   <2e-16 ***
plot5 150.1819     0.4524   332.0   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9047 on 15 degrees of freedom
Multiple R-squared:     1,      Adjusted R-squared:     1
F-statistic: 1.098e+05 on 5 and 15 DF,  p-value: < 2.2e-16
```

The residual s.e. for this model is 0.9047, which is similar to the one obtained previously. The residuals of this model are now centered around zero:



Both models we constructed so far were wrong, because they did not account for the fact that the plots we used were just a random sample from a large number of possible plots that could have been chosen. They did also not account for the pseudoreplication (several samples taken per plot).

This is evident if we look at the ANOVA table for model2:

```
> anova(model2)
Analysis of Variance Table

Response: yield
          Df Sum Sq Mean Sq F value    Pr(>F)
plot       5 449206   89841  109763 < 2.2e-16 ***
Residuals 15     12       1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We see that the estimates are based on a sample size of 20 data points, but there were only 5 plots in total.

One solution would be to analyse the experiment as a split-plot ANOVA. However, because there are no treatments applied below the plot scale, there are not enough degrees of freedom to test for significant differences, and we only get the corresponding sums of squares and variances:

```
model3=aov(yield~1+Error(plot))
summary(model3)
Error: plot
          Df  Sum Sq Mean Sq F value Pr(>F)
Residuals  4 1.16607 0.29152

Error: Within
          Df  Sum Sq Mean Sq F value Pr(>F)
Residuals 15 12.2775  0.8185
```

The only way to analyse these data in a sensible way is to use a **mixed effects model**. Suppose, for example, that we had 100 plots instead of 5; now the number of parameters in our “classical” linear models would increase linearly as more and more plots are added. The plots themselves, however, are “uninteresting” in the sense that we only want to predict mean plant yield and how much variance there is between plots. We are not interested in specific plot comparisons (for example “plot 33 differed significantly from plot 67”.

To understand the transition from fixed to mixed effects models, we first need to come back to our initial model formulation, which was (in this case)

model2: $y = \beta_1 X + \varepsilon$

```
model.matrix(model2)
plot1 plot2 plot3 plot4 plot5
1      1     0     0     0     0
2      1     0     0     0     0
3      1     0     0     0     0
```

```
4      1     0     0     0     0
5      0     1     0     0     0
6      0     1     0     0     0
7      0     1     0     0     0
8      0     1     0     0     0
(...)
contrasts(plot)
     2 3 4 5
1    0 0 0 0
2    1 0 0 0
3    0 1 0 0
4    0 0 1 0
5    0 0 0 1
```

You can see that four dummy variables have been introduced for the k-1 orthogonal contrasts of the factor "plot". This is completely not what we want! As we just said: We are not interested in those specific comparisons, because had there been 1000 plots, there would be 999 possible comparisons, and we would be very likely to find (just by chance alone) some plots differing significantly in "mean yield".

Hence, in mixed effects models, some or all of the parameters ß in a model are not treated as fixed parameters, but as random variables. This has the great advantage that it saves us a lot of degrees of freedom, and it allows an estimation of between-plot and within-plot variability.

Expressed as a mixed effects model, any linear model formula now becomes:

$classical\ linear\ model$: $y = ß_0 + \beta_1 x + \varepsilon$

$mixed\ model$: $y = ß_0 + b_0 + ß_1 x + b_1 x + \varepsilon$

Thus, there is now a mixture of both fixed effects ß, and random effects b. These random effects are now assumed to have mean 0 and variance sigma-squared.

Our model 1, expressed as a mixed-effects model, could now become

$$model\ 1: \qquad y = ß_0 + b_0 + \varepsilon$$

This means that a fixed intercept term $ß_0$ is estimated, but the deviations from this fixed effect are assumed to be random deviations between plots (b_0), plus random variation within plots (ε).

Let´s try this out in R:

```
library(nlme)
model4=lme(yield~1,random=~1|plot)
summary(model4)

Linear mixed-effects model fit by REML
 Data: NULL
       AIC      BIC    logLik
  56.34253 59.17585 -25.17127

Random effects:
 Formula: ~1 | plot
        (Intercept)  Residual
```

```
StdDev: 1.988913e-05 0.8411627

Fixed effects: yield ~ 1
               Value Std.Error DF  t-value p-value
(Intercept) 149.8674 0.1880897 15 796.7867       0

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-2.8804785 -0.2680754  0.0741240  0.6848041  1.9436069

Number of Observations: 20
Number of Groups: 5
```