Mixed models

November 20, 2019

Answers for this lab must be submitted before **December 4th at 5pm on Moodle**. In your answer for each question, please include a copy of the R code used (if applicable) and the results obtained.

1. Wheat yield as a function of fertilizer and soil moisture

The `Wheat` data frame included in the *nlme* package presents the result of an agricultural experiment to measure the effect of fertilizer quantity and soil moisture on the yield of a wheat variety (`DryMatter`, the response variable). The plants were divided into 12 trays. Soil moisture is constant in a tray, but each tray is divided into 4 sections each receiving a different amount of fertilizer. There are therefore 48 individual measurements of the yield, i.e. 4 per tray.

```
library(nlme)
data(Wheat)
head(Wheat)
```

```
## Grouped Data: DryMatter ~ fertilizer | Tray
##   Tray Moisture fertilizer DryMatter
## 1    1       10          2    3.3458
## 2    1       10          4    4.3170
## 3    1       10          6    4.5572
## 4    1       10          8    5.8794
## 5    2       10          2    4.0444
## 6    2       10          4    4.1413
```

(a) Estimate the parameters of a linear model including the additive effects of fertilizer and moisture on yield. Is it correct to ignore the grouping of the measurements per tray? Justify your answer from a plot of residuals.

(b) Now fit a mixed model that includes the same fixed effects, as well as a random effect representing the grouped data structure. Compare the coefficients of the fixed effects (fertilizer and moisture) for both models, as well as the standard errors of these coefficients. Explain the differences if any.

(c) From appropriate diagnostic plots, verify that the mixed model residuals in (b) are randomly distributed following a normal distribution.

(d) Calculate the intra-class correlation for the mixed model in (b). What is the mathematical interpretation of this value?

2. Growth curves of spruce trees

The `Spruce` data frame included in the *nlme* package contains data on the growth of 79 spruce trees. The logarithm of the volume (`logSize`) for each spruce (identified by the `Tree` number) was measured at 13 different times from the beginning of the experiment (`days`). The trees are also divided between 4 plots.

```
data(Spruce)
head(Spruce)
```

```
## Grouped Data: logSize ~ days | Tree
##    Tree days logSize plot
## 1 O1T01  152    4.51    1
```

```
## 2 O1T01  174    4.98    1
## 3 O1T01  201    5.41    1
## 4 O1T01  227    5.90    1
## 5 O1T01  258    6.15    1
## 6 O1T01  469    6.16    1
```

(a) Create a graph of the observed `logSize` for these trees, with a color code for the plot. Why would it be better to model the growth curve by taking `days` as a factor, i.e. `logSize ~ as.factor(days)` rather than `logSize ~ days`?

Tip: With `ggplot`, you can overlay a `geom_point` and `geom_line`, and then specify `group = Tree` in the `aes` function to ensure that lines are drawn between points for each tree.

(b) Estimate the parameters of the linear model `logSize ~ as.factor(days)`. According to this model, what is the average variation of `logSize` between day 152 and day 201? What is its standard error?

(c) Now fit a mixed model that takes into account that these are repeated measurements on the same trees. (*Note*: Ignore the plot for now.) Based on these results, why is it beneficial to measure the same trees on each sampling day to estimate a growth curve?

(d) What is the intra-class correlation for the model in (c)? Based on this result, is the variation in size among trees at any point in the experiment due more to (i) initial differences in size between trees or (ii) a change in the growth pattern from one tree to another?

(e) Add to the model in (c) a random effect for the plot. According to this model, does tree size differ significantly between plots?