Applied Statistics : Practical 11

This practical is about how to fit linear mixed effects models using the `lme4` package. You may need to install it first (using either the `install.packages` command or the menu bar). The dataset `biodiversity.txt` and `school.txt` can be found on the course webpage.

1. **Biodiversity data**
   In this item we analyse the dataset `biodiversity.txt`. It contains measurements of biodiversity taken in the different seasons in 5 different geographical locations. We are interested in investigating the variation in biodiversity between the seasons and we are going to treat the location as a random effect.

```
> eco_data<-read.table("biodiversity.txt",header=TRUE)
> attach(eco_data)
> library(lme4)
> eco_model<-lmer(Biodiversity~ season+(0+season|location),REML=FALSE)
> summary(eco_model)

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: Biodiversity ~ season + (0 + season | location)
##
##      AIC      BIC   logLik deviance df.resid
##    120.4    147.5    -45.2     90.4       30
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.14042 -0.69245 -0.00516  0.70756  1.85509
##
## Random effects:
##  Groups   Name         Variance Std.Dev. Corr
##  location seasonautumn 6.3422   2.5184
##           seasonspring 5.1691   2.2736   1.00
##           seasonsummer 5.3362   2.3100   1.00 1.00
##           seasonwinter 4.8388   2.1997   1.00 1.00 1.00
##  Residual              0.2425   0.4924
## Number of obs: 45, groups:  location, 5
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)   2.76863    1.13636   2.436
## seasonspring -0.06667    0.23897  -0.279
## seasonsummer  0.14374    0.23196   0.620
## seasonwinter  0.27825    0.25076   1.110
##
## Correlation of Fixed Effects:
##             (Intr) ssnspr ssnsmm
## seasonsprng -0.538
## seasonsummr -0.485  0.600
## seasonwintr -0.642  0.636  0.615
```

   Can you understand the model that have been fitted? Try to describe it, specifying which effects are fixed and which are random.

The model has a different (fixed) mean for the biodiversity for each season, since the factor `season` *is included in the fixed effects part of the model. Plus, the model has a random coefficient associated to the season for each location. This means that each location is allowed a different seasonal mean, whose difference with respect to overall seasonal mean is random.*

To decide if the fixed effect is significant in the model, we fit a model without the fixed effect and compare the likelihood ratio statistics with the χ^2 approximation.

```
> eco_model2<-lmer(Biodiversity˜ 1+(0+season|location),REML=FALSE)
> anova(eco_model2,eco_model)

## Data: NULL
## Models:
## eco_model2: Biodiversity ˜ 1 + (0 + season | location)
## eco_model: Biodiversity ˜ season + (0 + season | location)
##            Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## eco_model2 12 117.17 138.85 -46.585   93.170
## eco_model  15 120.42 147.52 -45.211   90.421 2.7483      3     0.4321
```

Which model is preferable?
The second model is indeed preferable, since the addition of the fixed term does not improve the likelihood significantly (the p-value of the likelihood ratio test is about 0.43*).*

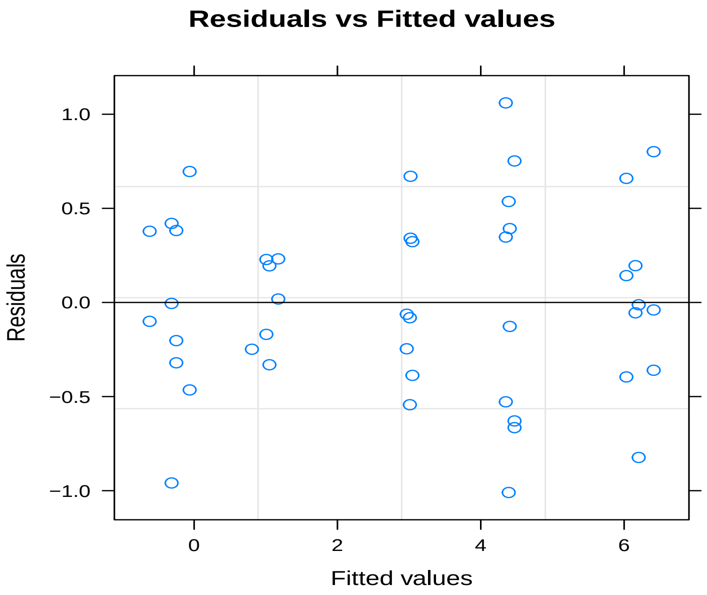
An alternative is of course to select the model using the AIC:

```
> AIC(eco_model)
> AIC(eco_model2)
```

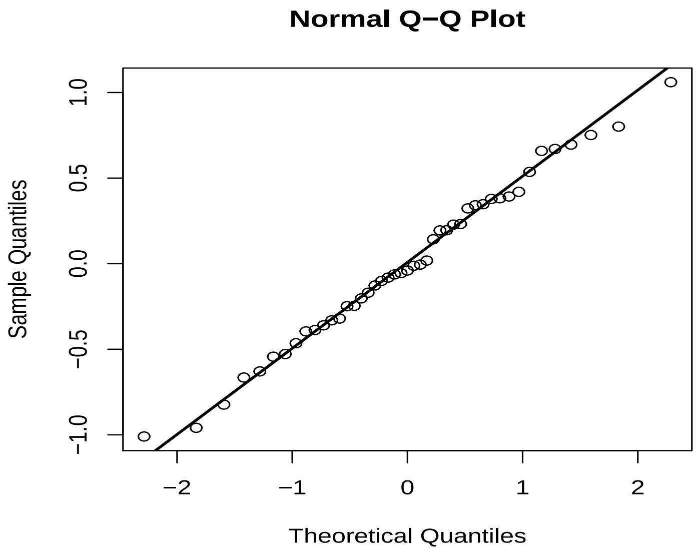
Which model is favored by the AIC?
The second model is again preferable because it has smaller AIC.

Diagnostics plots can be obtained as

```
> plot(eco_model2,xlab="Fitted values",ylab="Residuals",
+      main="Residuals vs Fitted values")
```



```
> qqnorm(residuals(eco_model2))
> qqline(residuals(eco_model2),lwd=2)
```



Can you see any problem with the model assumptions?
No.

2. **Nested effects**

When we have two random effects $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ and each level of $\boldsymbol{u}_1$ occurs with one and only one level of $\boldsymbol{u}_2$, we say that $\boldsymbol{u}_1$ is *nested* within $\boldsymbol{u}_2$. An example of this can be found in the dataset `Pastes` available in the `lme4` packages. This describes the measurements of `strength` of pastes coming from 3 different samples for each one of 10 different batches of material (the dataset contains also a `cask` variable that has been used to determine the `sample` information and you can ignore it).

```
> data(Pastes)
> head(Pastes)

##   strength batch cask sample
## 1     62.8     A    a    A:a
## 2     62.6     A    a    A:a
## 3     60.1     A    b    A:b
## 4     62.3     A    b    A:b
## 5     62.7     A    c    A:c
## 6     63.1     A    c    A:c

> attach(Pastes)
```

Here the sample is nested within the batch, since each instance of sampling may belong only to one specific batch in the experiment. This kind of situation appears frequently in practice and it can be fitted easily by including both random effects in the `lmer` command. Note that sometimes the effects are nested but the labels are not clear about this (for example, samples have been coded just with `a`,`b`,`c` for all batches) and you need to specify them in a way that `R` can interpret them as nested (see the last item of the practical for an example).

```
> pastes_model <- lmer(strength ~ 1 + (1|sample) + (1|batch), REML=TRUE)
```

Looking at the summary of the model, which are the variances of the two random effects? (Note that for the way we have defined the model, the random effects are uncorrelated, i.e. we have constrained G to be diagonal).
The variance associated to the sample is 8.434 *and the variance associated to the batch is* 1.657*.*

Discuss the proportion of the two variances, does this suggest to fit a simpler model?
Since the variance associated to the batch is far smaller than the one associated to the sample, this may suggest to fit a model without the batch effect.

We want now to check if the batches effect is really needed in the model, i.e. if there is a difference in the strength due to the different batches or all the variability is due to the sample variability. Remember that the χ^2 approximation for the likelihood ratio statistics is untrustworthy for the random effects. To test for the random effects, we prefer to use a parametric bootstrap. This asks for some work on our part. First, we need to fit the null model in the comparison, in this case it is

```
> pastes_model0<-lmer(strength~ 1+(1|sample))
```

Then, we simulate 100 times (1000 would be better, but let us save some time here) from this model and for each time we compute the likelihood ratio (and save the result) between the fit of the two models:

```
> likelihood_ratio<-rep(NA,100)
> for (k in 1:100){
+   y<-simulate(pastes_model0)[[1]]
+   boot_null<-lmer(y~ 1+(1|sample),REML=TRUE)
+   boot_model2<-lmer(y~ 1+(1|sample)+(1|batch),REML=TRUE)
+   likelihood_ratio[k]<- as.numeric(2*(logLik(boot_model2)- logLik(boot_null)))
+ }
```

Finally, the p-value approximated by the parametric bootstrap is the proportion of simulations for which the likelihood ratio statistics is larger than the one in the original sample.

```
> mean(likelihood_ratio>as.numeric(2*(logLik(pastes_model)- logLik(pastes_model0))))
```

What can you conclude about the batch effect?
The batch effect is not needed in the model.

Which model would be selected by the AIC?

```
> AIC(pastes_model,pastes_model0)

##               df      AIC
## pastes_model   4 254.9907
## pastes_model0  3 253.6484
```

The simpler model would be selected, because it has smaller AIC.

3. **Sleepstudy data**
   The dataset `sleepstudy` in the package `lme4` contains measurements of reaction times after a certain amount of days of sleep deprivation for 18 different subjects.

```
> data("sleepstudy")
> 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

> attach(sleepstudy)
```

We can plot the data in a useful way using the function `xyplot` in the package `lattice`:

```
> library(lattice)
> xyplot(Reaction~Days|Subject)
```

This plots the pairs of days of sleep deprivation and reaction times in a separate panel for each subject. Comment on this plot, which type of model is suggested?
It suggests that there is a (roughly) linear relationship between the number of days and the reaction time but both the intercept and the slope are subject-dependent. We could therefore propose a model with both intercept and slope as random effects associated to the subject.

Consider now the following two models:

```
> sleep_model1<-lmer(Reaction ~ Days + (Days|Subject),REML=FALSE)
> sleep_model2<-lmer(Reaction ~ Days + (1|Subject) +(0+Days|Subject),REML=FALSE)
```

Looking at the summary of the two models and the examples seen in the previous items, can you understand what is the difference between them?
The second model constrains the random effects associated to the intercept and the slope to be uncorrelated, since they are defined in different parenthesis. This can be seen from the summary, because the output for the second model does not contain an estimate for the correlation.

(An alternative way to fit the second model is

```
> sleep_model2<-lmer(Reaction ~ Days + (Days||Subject), REML=FALSE)
```

) Which model is better?
The second model has a smaller AIC and it is therefore preferable. However, the diagnostics plots show evidence of the presence of outliers and this may question the validity of the model.

Is the fixed effect needed in the model? What does this imply for the relationship between the days of sleep deprivation and the reaction time for the general population?

Yes, the p-values of the likelihood ratio test between the model with and without the fixed effect associated to the number of day is very small. We can conclude that, on average *in the population, the reaction time increases with the number of days with sleep deprivation*

It is also possible to use the conditional means of the random effects (provided by the `ranef` function) to fit a subject-specific model. The command

```
> u_hat<-ranef(sleep_model2)
```

returns the conditional means for the random effects for all the subjects. Try to fit the model you have chosen for the subject coded as 308 and superimpose the regression line to the data. (Note that `u_hat` is a list with one element for each grouping variable, in this case just one, `Subject`, and you need to access it as `u_hat$Subject`)

```
> subject<-which(Subject==308)
> plot(Days[subject],Reaction[subject])
> points(Days[subject],251.405+10.467*Days[subject]
+        +u_hat$Subject[1,1]+u_hat$Subject[1,2]*Days[subject],type='l',col=4)
```

4. **School data** The data set `school.txt` contains an amended version of the data available from Jon Starkweather webpage at `http://www.unt.edu/rss/class/Jon/R_SC/Module9/lmm.data.txt`. Measurements for openness, agreeableness, social ability and extroversion are provided for children belonging to different schools and different classes within each school. This means that we want to treat class and school as nested random effects.

```
> school_data<-read.table("school.txt",header=TRUE, sep=",")
> attach(school_data)
> head(school_data)

##   id    extro     open    agree    social class school
## 1  1 63.69356 43.43306 38.02668  75.05811     d     IV
## 2  2 69.48244 46.86979 31.48957  98.12560     a     VI
## 3  3 79.74006 32.27013 40.20866 116.33897     d     VI
## 4  4 62.96674 44.40790 30.50866  90.46888     c     IV
## 5  5 64.24582 36.86337 37.43949  98.51873     d     IV
## 6  6 50.97107 46.25627 38.83196  75.21992     d      I
```

Note however that the classes are labelled as a,b, c and d for every school and this means that `R` cannot understand that the two effects are nested. We need first to define a new variable that contains their interaction:

```
> schoolClass<-class:school
```

This new variable has a different level for each combination class-school and it will be treated as nested into the school variable.

Now choose the best model to explain the relationship between the extroversion and social ability.You can start from

```
> school_model1<-lmer(extro ~ social + (social|school)+(0+social|schoolClass))
```

The variance of the slope for the school-level random effects is very small and we can try to remove it from the model

```
> school_model2<-lmer(extro ~ social + (1|school)+(0+social|schoolClass))
> AIC(school_model1,school_model2)

##               df      AIC
## school_model1  7 3763.904
## school_model2  5 3759.907
```

This results in a decrease of the AIC and we therefore prefer the second model. Let us now try to fit a model without random slopes.

```
> school_model3<-lmer(extro ~ social + (1|school))
> AIC(school_model3,school_model2)

##               df      AIC
## school_model3  4 5822.639
## school_model2  5 3759.907
```

However, the new model has a much larger AIC, thus we keep the random class-associated slope in the model. Finally, let us check if we can remove the fixed effect from the model

```
> school_model4<-lmer(extro ~  (1|school),REML=FALSE)
> anova(school_model4,school_model2)

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

## Data: NULL
## Models:
## school_model4: extro ~ (1 | school)
## school_model2: extro ~ social + (1 | school) + (0 + social | schoolClass)
##               Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## school_model4  3 5816.6 5831.8 -2905.3   5810.6
## school_model2  5 3755.8 3781.3 -1872.9   3745.8 2064.7      2  < 2.2e-16
##
## school_model4
## school_model2 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

and we find that we need to keep the fixed slope in the model. We conclude that, on average, larger social ability leads to a smaller measure of extroversion but there is variability at the class level.