Practical 4: Random effects model

Verena Zuber, Jelena Besevic, Alpha Forna, and Saredo Said

14/2/2019

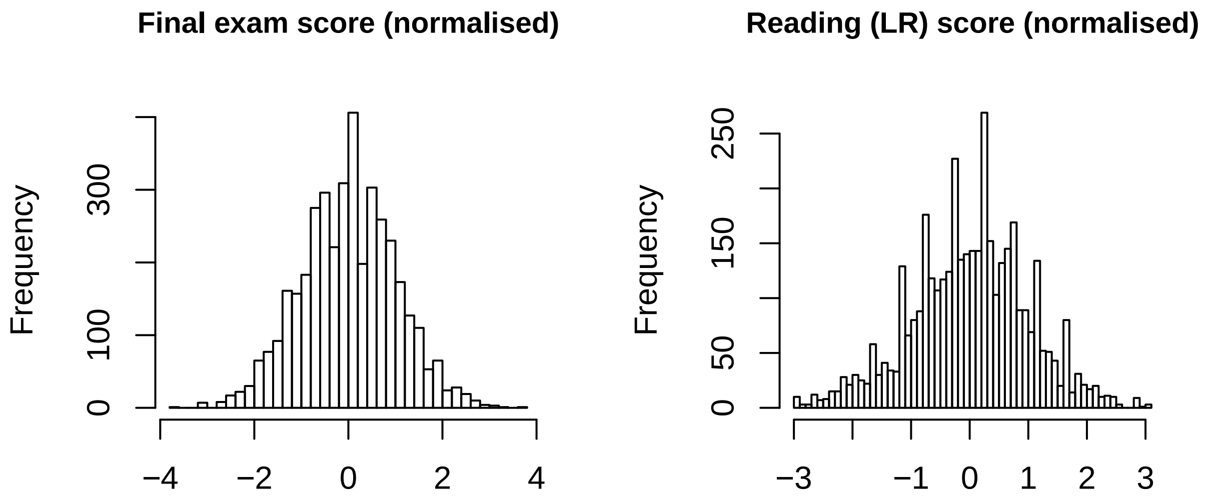
Part 1: Linear mixed model: Exam scores from London

The first part of the practical considers exam scores of 3,935 students from 65 schools in Inner London. In particular, we want to find out how the final exam score can be predicted by reading abilities as measured in the London reading (LR) test.

```
load("exam.London")
dim(exam)
```

```
## [1] 3935   10
```

```
par(mfrow=c(1,2))
hist(exam$normexam, breaks =50, main="Final exam score (normalised)", cex.main=0.9, xlab="")
hist(exam$standLRT, breaks =50, main="Reading (LR) score (normalised)", cex.main =0.9, xlab="")
```



Additional covariates of the data are:

- school: School ID - a factor
- schgend: School gender - a factor. Levels are 'mixed', 'boys', and 'girls'
- schavg: School average of intake score
- vr: Student level Verbal Reasoning (VR) score band at intake - 'bottom 25%', 'mid 50%', and 'top 25%'
- intake: Band of student's intake score - 'bottom 25%', 'mid 50%' and 'top 25%'
- sex: Sex of the student - levels are 'F' and 'M'
- type: School type - levels are 'Mxd' and 'Sngl'
- student: Student id (within school) - a factor

Question 1.1

Fit a linear model to test if there is a linear relationship between reading ability and the final exam score and plot a scatterplot of exam score against reading ability.

Question 1.2

Are there any potential issues with the standard linear model?

Question 1.3

Fit a fixed effect model accounting for the effect of schools using the lm() function where you add school (as.factor()) as covariate. What is the interpretation of the model and how many additional parameters do we need to estimate?

Question 1.4

Now use the function in the lme function in the

```
library(nlme)
```

package to estimate a random effects model with a random intercept depending on the school. What is the interpretation of the fixed effect? How many parameters do we need to estimate compared to the fixed effects model?

Question 1.5

What is the intra-class correlation coefficient for this model (lecture 4, slide 39) and how do you interpret it?

Question 1.6

Add a random slope depending on school to your model and see if the effect of the fixed effects changes.

Question 1.7

Which of the covariates are individual-level and which are group-level variables? Re-fit your random intercept model adding the group-level variables to the random effects model.

Question 1.8

Compare the random intercept (Q1.4) and the random intercept and slope model (Q1.5) using the likelihood ratio test and discuss which one has the better model fit.

Question 1.9

Compare the random intercept (Q1.4) and the one with the additional covariate (Q1.6) using the AIC and BIC (note that those two models are not nested) and discuss which one has the better model fit.

Part 2: Linear mixed model: Survival on the Titanic

The sinking of the titanic was one of the greatest disaster in navel history. After colliding with an iceberg, the titanic sank and 1,502 out of 2,224 passengers and crew were killed. The following data set has collected information on n=1,309 of the passengers and their survival.

```
titanic = read.csv("titanic.csv")
dim(titanic)
```

```
## [1] 1309   14
```

```
table(titanic$survived)
```

```
##
##   0   1
## 809 500
```

The dataset includes:

- survival: Survival (0 = No; 1 = Yes)
- class: Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
- name: Name
- sex: Sex (1=female, 2=male)
- age: Age
- sibsp: Number of Siblings/Spouses Aboard

- parch: Number of Parents/Children Aboard
- ticket: Ticket Number
- fare: Passenger Fare
- cabin: Cabin
- embarked: Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)
- boat: Lifeboat (if survived)
- body: Body number (if did not survive and body was recovered)

For more information on the data and a data challenge called ‘Machine Learning from Disaster’ see

https://www.kaggle.com/c/titanic

In the following we want to test if the phrase ‘women and children first’ was adapted for the evacuation of the titanic.

Question 2.1

Since survival is a binary outcome here, use a glm to test if age and sex had an effect on survival.

Question 2.2

Next step is to account for the passenger class (variable pclass) in a fixed effects model and discuss the implications and difference to the simple model.

Question 2.3

Discuss whether to include the passenger class as a fixed or random effect and fit a random effects model with a random intercept depending on passenger class using the glmer() function in the

```r
library(lme4)
```

package.

Question 2.4

Add a random slope depending on passenger class to your model and compare it with the random intercept only model using a likelihood test.

Question 2.5

How do you explain the difference in results after accounting for passenger class? Use a boxplot and a violin plot for age depending on passenger class to illustrate your argument.

Part 3 (optional): Decision trees and random forests: Survival on the Titanic

The final part of this practical uses decision trees and random forest to analyse the titanic data. Make sure to have the following two packages

```r
library(tree)
library(randomForest)
```

installed.

Question 3.1

Fit a decision tree on the titanic data using the following predictor matrix including passenger class, sex, age, number of siblings/spouses aboard, and number of parents/children aboard after excluding missing values.

```r
x=cbind(titanic$pclass, titanic$sex, titanic$age, titanic$sibsp, titanic$parch)
rm = which(is.na(titanic$age)==TRUE)
x.input = x[-rm,]
dim(x.input)
```

```
## [1] 1046    5
```

```
colnames(x.input) = c("pclass", "sex", "age", "sibsp", "parch")
y.input = as.factor(titanic$survived[-rm])
table(y.input)
```

```
## y.input
##   0   1
## 619 427
```

Use the function tree in the tree package.

Question 3.2

What is a concern when fitting a single decision tree?

Question 3.3

Prune your tree using cross-validation (cv.tree) and use the option FUN = prune.misclass for the misclassification rate as criterion. Choose the model with the lowest misclassification error and plot the tree. How do you interpret the decision tree?

Question 3.4

Finally fit a random forest to the data and look at the variable importance. What was the key variable for survival in the titanic disaster?