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.

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

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

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

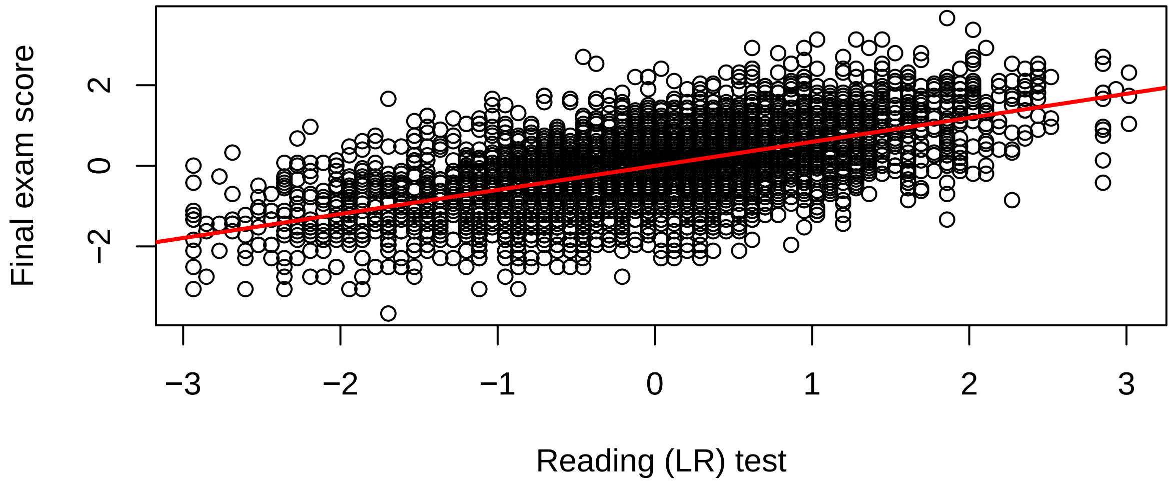
Question 1.1

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.

```r
exam.lm = lm(normexam~standLRT, data=exam)
summary(exam.lm)
```

```
##
## Call:
## lm(formula = normexam ~ standLRT, data = exam)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6518 -0.5182  0.0122  0.5404  2.9766
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.003163   0.012873  -0.246    0.806
## standLRT     0.596474   0.012972  45.981   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8075 on 3933 degrees of freedom
## Multiple R-squared:  0.3496, Adjusted R-squared:  0.3495
## F-statistic:  2114 on 1 and 3933 DF,  p-value: < 2.2e-16
```

```r
plot(exam$standLRT,exam$normexam, xlab="Reading (LR) test", ylab="Final exam score")
abline(exam.lm, lwd=2, col="red")
```



Reply: There is a linear relationship between reading ability (standardised London reading test) and the final exam score. An increase of 1 in the reading score, increases the final exam score by 0.596474.

Question 1.2

Are there any potential issues with the standard linear model?

Reply: Yes, this linear model treats all observations as independent and disregards the potential group structure as induced by the school in which the students are studying.

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?

```
exam.fe = lm(normexam~standLRT+as.factor(school), data=exam)
coef(exam.fe)
```

```
##         (Intercept)            standLRT  as.factor(school)2
##         0.405549804         0.560181196         0.156197429
##  as.factor(school)3  as.factor(school)4  as.factor(school)5
##         0.168658362        -0.383878052        -0.134037654
##  as.factor(school)6  as.factor(school)7  as.factor(school)8
##         0.181827272        -0.026970604        -0.413833357
##  as.factor(school)9 as.factor(school)10 as.factor(school)11
##        -0.523084933        -0.780966981        -0.179456700
## as.factor(school)12 as.factor(school)13 as.factor(school)14
##        -0.483312036        -0.571058377        -0.581067253
## as.factor(school)15 as.factor(school)16 as.factor(school)17
##        -0.606702657        -0.825202132        -0.595818835
## as.factor(school)18 as.factor(school)19 as.factor(school)20
##        -0.490771143        -0.431115211        -0.205049233
## as.factor(school)21 as.factor(school)22 as.factor(school)23
##        -0.126419704        -0.880185131        -1.000512830
## as.factor(school)24 as.factor(school)25 as.factor(school)26
##        -0.179747660        -0.666833203        -0.429359943
## as.factor(school)27 as.factor(school)28 as.factor(school)29
##        -0.378424526        -1.106500587        -0.163003687
## as.factor(school)30 as.factor(school)31 as.factor(school)32
##        -0.212081930        -0.414259745        -0.411342710
## as.factor(school)33 as.factor(school)34 as.factor(school)35
##        -0.356107503        -0.537174088        -0.214086960
## as.factor(school)36 as.factor(school)37 as.factor(school)38
##        -0.640611796        -0.647423110        -0.574310199
## as.factor(school)39 as.factor(school)40 as.factor(school)41
##        -0.241066377        -0.645538637        -0.170886794
## as.factor(school)42 as.factor(school)43 as.factor(school)44
##        -0.306591722        -0.497881063        -0.688387983
## as.factor(school)45 as.factor(school)46 as.factor(school)47
##        -0.520295271        -0.788708334        -0.455262007
## as.factor(school)48 as.factor(school)49 as.factor(school)50
##        -0.587882734        -0.370351439        -0.744855302
## as.factor(school)51 as.factor(school)52 as.factor(school)53
##        -0.461041141         0.008072604         0.398475436
## as.factor(school)54 as.factor(school)55 as.factor(school)56
##        -1.368601638         0.184024334        -0.427019374
## as.factor(school)57 as.factor(school)58 as.factor(school)59
##        -0.377376447        -0.241451121        -1.128811022
## as.factor(school)60 as.factor(school)61 as.factor(school)62
##        -0.143876751        -0.438222462        -0.493550060
## as.factor(school)63 as.factor(school)64 as.factor(school)65
##         0.242290425        -0.304716265        -0.593671312
```

Reply: Also the fixed effects model finds a linear relationship where the regression coefficient equals 0.560181196. By introducing the factor school as covariate we need to estimate additional to the intercept and the beta regression coefficient another 64 regression parameter, one for each school minus the reference category. In total these are 66 parameters to estimate in a fixed effects model.

Question 1.4

Now use the function in the lme function in the

```r
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?

```r
RandomIntercept = lme(normexam~standLRT,  random = ~1|school, data=exam)
summary(RandomIntercept)
```

```
## Linear mixed-effects model fit by REML
##  Data: exam
##        AIC      BIC    logLik
##   9107.902 9133.011 -4549.951
##
## Random effects:
##  Formula: ~1 | school
##         (Intercept)  Residual
## StdDev:   0.3071927 0.7535887
##
## Fixed effects: normexam ~ standLRT
##                 Value  Std.Error   DF  t-value p-value
## (Intercept) 0.0003680 0.04052258 3869  0.00908  0.9928
## standLRT    0.5641318 0.01269508 3869 44.43704  0.0000
##  Correlation:
##          (Intr)
## standLRT 0.007
```

```
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -3.69766166 -0.62537566  0.02378645  0.67363330  3.25607135
##
## Number of Observations: 3935
## Number of Groups: 65
```

Reply: There is almost no change in the interpretation of the random effects model, also here the reading ability has a linear relationship with the exam score (beta=0.5641318). We need to estimate two additional parameters, StdDev(Intercept), StdDev(Residual), so in total there are 4 parameters, 62 parameters less than in 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?

```
std.u = 0.3071927
std.e = 0.7535887
rho = std.u^2 / (std.u^2+std.e^2)
rho
```

```
## [1] 0.1424922
```

Reply: The intra-class correlation coefficient is 0.1424922, which is not strong, but still it should be accounted for.

Question 1.6

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

```
RandomSlope = lme(fixed=normexam~standLRT, random = ~ 1 + standLRT | school, data = exam)
summary(RandomSlope)
```

```
## Linear mixed-effects model fit by REML
##  Data: exam
##        AIC      BIC    logLik
##   9073.433 9111.096 -4530.716
##
## Random effects:
##  Formula: ~1 + standLRT | school
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev    Corr
## (Intercept) 0.3036509 (Intr)
## standLRT    0.1195903 0.515
## Residual    0.7457557
##
## Fixed effects: normexam ~ standLRT
##                  Value  Std.Error   DF   t-value p-value
## (Intercept) -0.0128591 0.04019262 3869 -0.319937   0.749
## standLRT     0.5572344 0.01999325 3869 27.871125   0.000
##  Correlation:
##          (Intr)
## standLRT 0.374
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -3.81039328 -0.63401763  0.03276076  0.67303433  3.44692923
##
```

```
## Number of Observations: 3935
## Number of Groups: 65
```

Reply: No, the strength of the fixed effect of reading ability stays roughly constant with a beta equal to 0.5572344.

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.

```
RandomInterceptCov = lme( normexam~standLRT + schavg + schgend, random = ~ 1 | school, data = exam)
summary(RandomInterceptCov)
```

```
## Linear mixed-effects model fit by REML
##  Data: exam
##        AIC     BIC    logLik
##   9104.396 9148.33 -4545.198
##
## Random effects:
##  Formula: ~1 | school
##         (Intercept)  Residual
## StdDev:   0.2649784 0.7537166
##
## Fixed effects: normexam ~ standLRT + schavg + schgend
##                   Value  Std.Error   DF  t-value p-value
## (Intercept)  -0.0814454 0.04847661 3869 -1.68010  0.0930
## standLRT      0.5601841 0.01276782 3869 43.87470  0.0000
## schavg        0.3519672 0.10711677   61  3.28583  0.0017
## schgendboys   0.1158202 0.10330517   61  1.12115  0.2666
## schgendgirls  0.2404062 0.08041229   61  2.98967  0.0040
##  Correlation:
##              (Intr) stnLRT schavg schgndb
## standLRT      0.000
## schavg        0.053 -0.119
## schgendboys  -0.465 -0.001  0.053
## schgendgirls -0.603 -0.001 -0.031  0.280
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -3.69621679 -0.62854698  0.02257168  0.67400518  3.22267911
##
## Number of Observations: 3935
## Number of Groups: 65
```

Reply: Group-level covariates are school gender, school average of intake score and school type. We add here the covariates school gender and school average of intake score. We find that the school average of intake score and the effect of being in a girls only school have a significant positive effect. The fixed effect of reading ability (beta=0.5601841) is comparable to other models.

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.

```
anova(RandomIntercept, RandomSlope)
```

```
##                 Model df      AIC      BIC    logLik   Test  L.Ratio
## RandomIntercept     1  4 9107.902 9133.011 -4549.951
## RandomSlope         2  6 9073.433 9111.096 -4530.716 1 vs 2 38.46905
##                 p-value
## RandomIntercept
## RandomSlope      <.0001
```

Reply: The random slope model has a better model fit than the random intercept model in all criteria, likelihood ratio (these are nested models, so it is ok to interpret the the likelihood ratio test here), AIC and BIC.

Question 1.9

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

```
anova(RandomIntercept, RandomInterceptCov)
```

```
##                    Model df      AIC      BIC    logLik   Test  L.Ratio
## RandomIntercept        1  4 9107.902 9133.011 -4549.951
## RandomInterceptCov     2  7 9104.396 9148.330 -4545.198 1 vs 2 9.506363
##                    p-value
## RandomIntercept
## RandomInterceptCov  0.0233
```

Reply: Since the two models are not nested we need to compare them using the AIC and BIC. We find that there is no conclusive improvement when adding the covariates. Using the AIC the covariate model would be better, using the BIC the intercept only model is better.

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.

```
glm.out = glm(as.factor(survived)~as.factor(sex)+age, family = "binomial", data = titanic)
summary(glm.out)
```

```
##
## Call:
## glm(formula = as.factor(survived) ~ as.factor(sex) + age, family = "binomial",
##     data = titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7247  -0.6859  -0.6603   0.7555   1.8737
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.235414   0.192032   6.433 1.25e-10 ***
## as.factor(sex)male -2.460689   0.152315 -16.155  < 2e-16 ***
## age                -0.004254   0.005207  -0.817    0.414
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1414.6  on 1045  degrees of freedom
## Residual deviance: 1101.3  on 1043  degrees of freedom
##   (263 observations deleted due to missingness)
## AIC: 1107.3
##
## Number of Fisher Scoring iterations: 4
```

Reply: We indeed see a worse chance of survival for men, but no effect of age.

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.

```
glm.out.fixed = glm(as.factor(survived)~as.factor(pclass)+as.factor(sex)+age,
 family = "binomial", data = titanic)
summary(glm.out.fixed)
```

```
##
## Call:
## glm(formula = as.factor(survived) ~ as.factor(pclass) + as.factor(sex) +
##     age, family = "binomial", data = titanic)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -2.6399  -0.6979  -0.4336   0.6688   2.3964
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         3.522074   0.326702  10.781  < 2e-16 ***
## as.factor(pclass)2 -1.280570   0.225538  -5.678 1.36e-08 ***
## as.factor(pclass)3 -2.289661   0.225802 -10.140  < 2e-16 ***
## as.factor(sex)male -2.497845   0.166037 -15.044  < 2e-16 ***
## age                -0.034393   0.006331  -5.433 5.56e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1414.62  on 1045  degrees of freedom
## Residual deviance:  982.45  on 1041  degrees of freedom
##   (263 observations deleted due to missingness)
## AIC: 992.45
##
## Number of Fisher Scoring iterations: 4
```

Reply: After accounting for passenger class, we find a significant effect of age. The beta coefficient is negative, which means that there is a better chance for survival for younger passengers. This indicates that we need to take into account the class, in order to see the true effect of age. If we do not account for the passenger class, this might act as a confounder. In the fixed effect we can also interpret the effect of the passenger class, in particular the best survival is seen for the first class, the reference category. The worst chance of survival is for passengers in the third class.

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.

```r
lmm.intercept = glmer(as.factor(survived)~as.factor(sex)+age+(1|pclass),
  family = "binomial", data = titanic)
summary(lmm.intercept)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: as.factor(survived) ~ as.factor(sex) + age + (1 | pclass)
##    Data: titanic
##
##      AIC      BIC   logLik deviance df.resid
##   1004.9   1024.7   -498.4    996.9     1042
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.4491 -0.5260 -0.3182  0.5015  4.0059
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  pclass (Intercept) 0.8512   0.9226
## Number of obs: 1046, groups:  pclass, 3
##
## Fixed effects:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         2.301841   0.583869   3.942 8.07e-05 ***
## as.factor(sex)male -2.496596   0.165643 -15.072  < 2e-16 ***
## age                -0.033552   0.006332  -5.299 1.16e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) as.f()
## as.fctr(sx) -0.195
## age         -0.344  0.065
```

Reply: When fitting a random effects model we do account for potential group structure induced by the passenger class, but we do not estimate the effect of each passenger class per se. There is no easy formula for the intra-class correlation coefficient as for linear outcome. We would need to call a different package (sjstats::icc and package glmmTMB to compute the generalised linear mixed model) to compute it. If we are only interested in the hypothesis 'Women and children first' we can evaluate this using the generalised mixed model. The interpretation of fixed and random effects with respect to the hypothesis is the same, both support an effect of sex and age on survival. Since class has only three categories it would be fine to use the fixed effects model (adding two additional parameters to the model). For comparison, the random intercept model has one additional parameter. In the fixed effects model we can additionally interpret the class specific survival chance.

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.

```
lmm.slope = glmer(as.factor(survived)~as.factor(sex)+age+(age|pclass),
 family = "binomial", data = titanic)
summary(lmm.slope)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: as.factor(survived) ~ as.factor(sex) + age + (age | pclass)
##    Data: titanic
##
##      AIC      BIC   logLik deviance df.resid
##   1008.5   1038.3   -498.3    996.5     1040
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.2804 -0.5201 -0.3173  0.4922  3.8245
##
## Random effects:
##  Groups Name        Variance  Std.Dev. Corr
##  pclass (Intercept) 9.276e-01 0.963141
##         age         7.439e-05 0.008625 -0.35
## Number of obs: 1046, groups:  pclass, 3
##
## Fixed effects:
```

```
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           2.328341   0.607940   3.830 0.000128 ***
## as.factor(sex)male -2.506048   0.166749 -15.029  < 2e-16 ***
## age                 -0.034232   0.008161  -4.194 2.74e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##               (Intr) as.f()
## as.fctr(sx) -0.194
## age         -0.460  0.065
```

Reply: Also the random slope and intercept model supports an effect of both sex and age.

```
anova(lmm.intercept, lmm.slope)
```

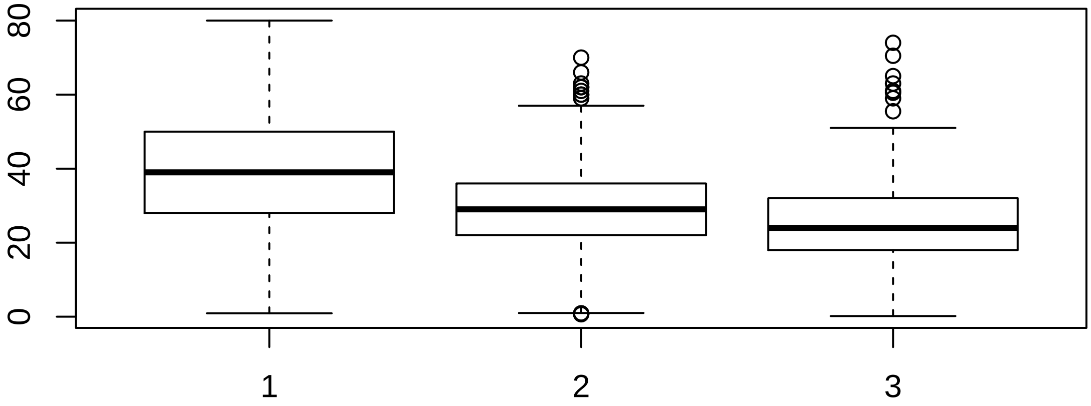
```
## Data: titanic
## Models:
## lmm.intercept: as.factor(survived) ~ as.factor(sex) + age + (1 | pclass)
## lmm.slope: as.factor(survived) ~ as.factor(sex) + age + (age | pclass)
##                Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## lmm.intercept  4 1004.9 1024.7 -498.45   996.90
## lmm.slope      6 1008.5 1038.3 -498.27   996.54 0.3555      2     0.8371
```

Reply: When considering random effects model, the random intercept model has a better model fit.

Question 2.5

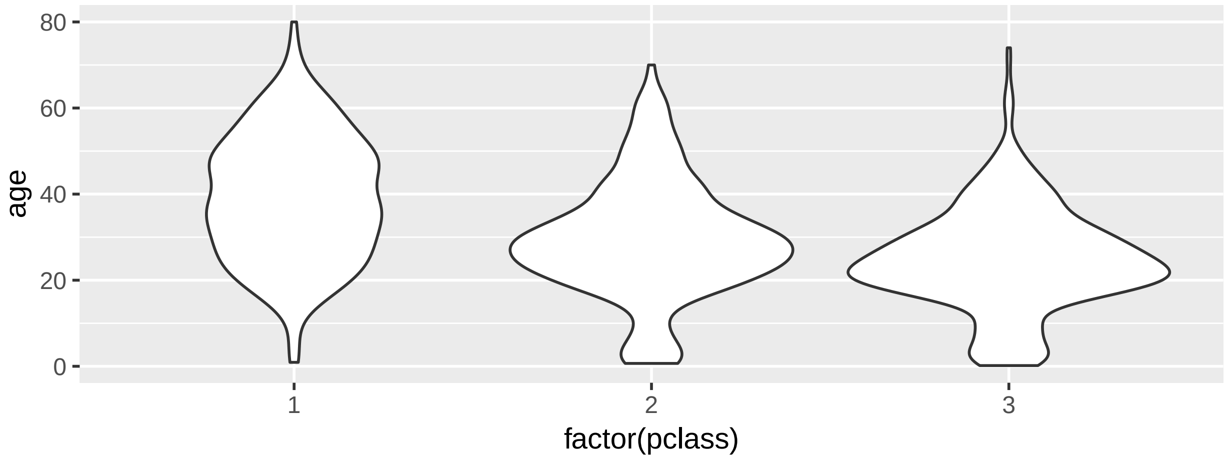
How do you explain the difference in results after accounting for passenger class?

```
boxplot(titanic$age~titanic$pclass)
```



Reply: There is a negative correlation between passenger class and age. Passengers travelling in lower classes were younger than passengers in the first class as shown in the boxplot of age grouped by passenger class.

```
library(ggplot2)
p = ggplot(titanic, aes(factor(pclass), age))
p + geom_violin(na.rm = TRUE)
```



As an aside: An alternative for the boxplot is the violin plot, a mixture of a boxplot and a kernel density function. Here it is even more obvious that there were much more young children travelling in class 2 and 3. Especially young children had a good chance for survival. In order to estimate the true effect of age we need to adjust for passenger class (either in a fixed or random effects model). Otherwise the estimate would be confounded since there where few children travelling first class.

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

```
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.

```
x=cbind(titanic$pclass, titanic$sex, titanic$age, titanic$sibsp, titanic$parch)
rm = which(is.na(titanic$age)==TRUE)
#alternatively use rm = which(!is.na(titanic$age)==TRUE) or rm = which(is.na(titanic$age)==FALSE)
x.input = x[-rm,]
dim(x.input)
```

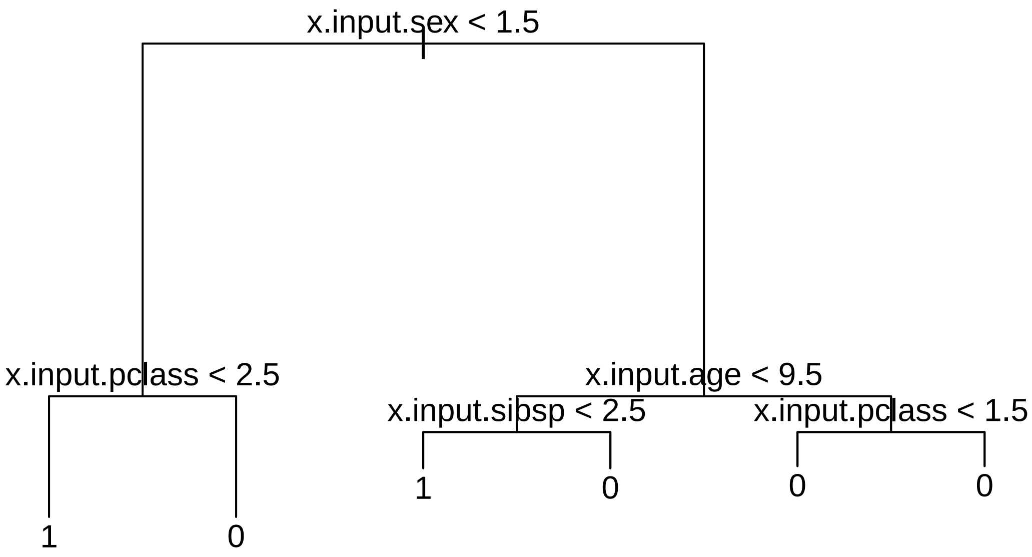
```
## [1] 1046    5
```

```
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.

```
tree.out = tree(y.input ~ x.input)
plot(tree.out)
text(tree.out)
```



Question 3.2

What is a concern when fitting a single decision tree?

Reply: Decision trees are prone to over-fitting. This can be prevented by either cross-validation or performing a random forest approach.

Question 3.3

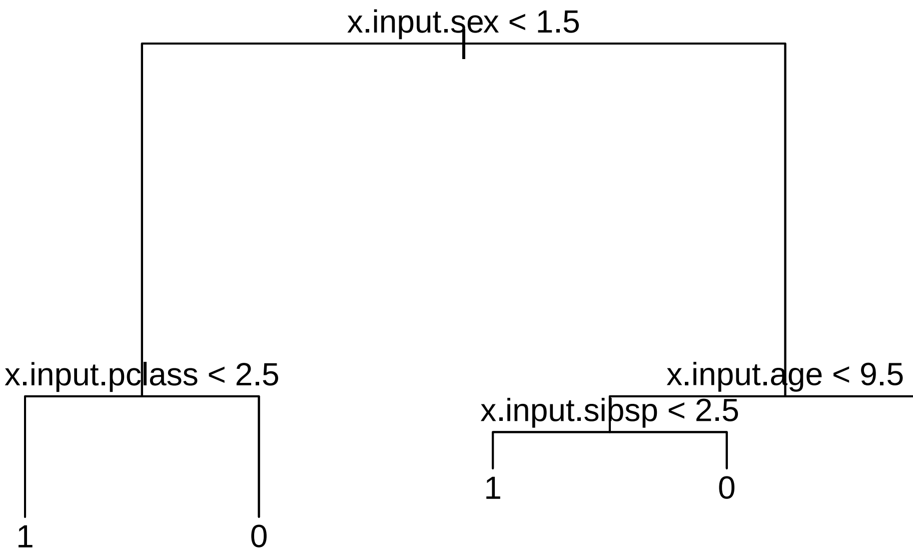
Prune your tree using cross-validation (cv.tree) and use the option FUN = prune.misclass for the misclassification rate as criterion. Choose the model with the lowest misclassification error and plot the tree. How do you interpret the decision tree?

```
set.seed(33)
cv.out = cv.tree(tree.out, FUN=prune.misclass)
cv.out
```

```
## $size
## [1] 6 5 4 2 1
##
## $dev
## [1] 203 203 217 231 427
##
## $k
## [1]  -Inf   0.0   8.0  10.5 196.0
##
## $method
## [1] "misclass"
##
## attr(,"class")
## [1] "prune"         "tree.sequence"
```

Reply: The models with size 6 and 5 have the same missclassification error ($dev=203). We decide to use the smaller model of size 5.

```
pruned.tree = prune.tree(tree.out, best=5)
plot(pruned.tree)
text(pruned.tree)
```

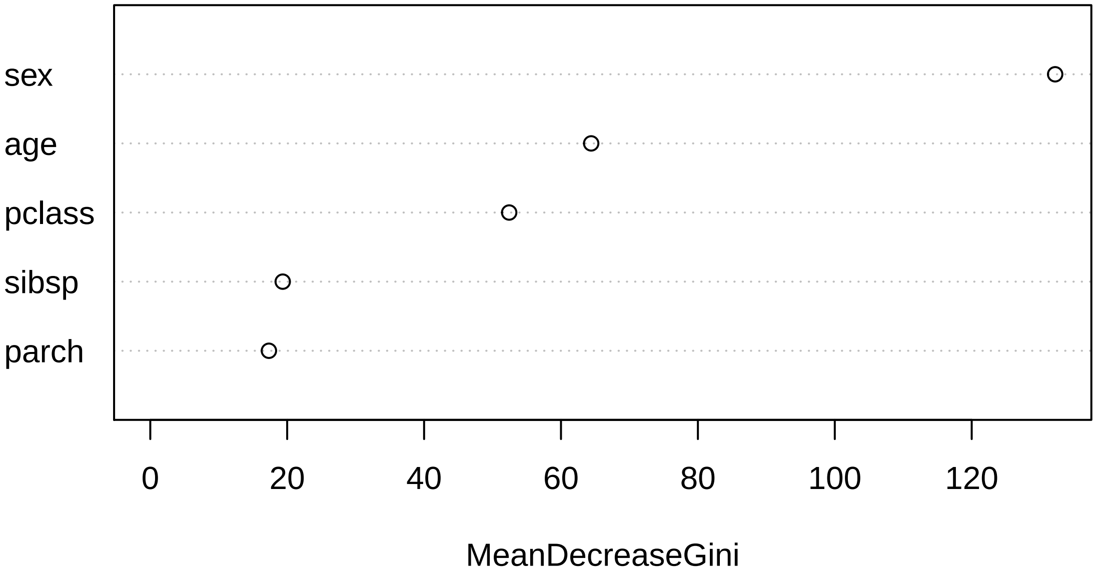


Interpretation: The first split is for female on the left and male on the right. For female the next split is passenger class, where class 1 and 2 are predicted to survive. For male the next split is on age, where men older than 9.5 years do not survive and men younger than 9.5 have another split at the variable siblings. Boys with fewer than 3 siblings survived, while boys with more siblings do not survive.

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?

```
rf.out = randomForest(y=y.input, x=x.input)
varImpPlot(rf.out, main="")
```



Reply: The most important variables are sex and age, reinforcing the 'women and children first' hypothesis. Third important variable is passenger class. The other two variables, sibsp (Number of Siblings/Spouses Aboard) and parch (Number of Parents/Children Aboard) seem to have little effect compared to the other variables.