EDDA - Assignment 3 - Group 77

Dante de Lang, Ignas Krikštaponis and Kamiel Gülpen

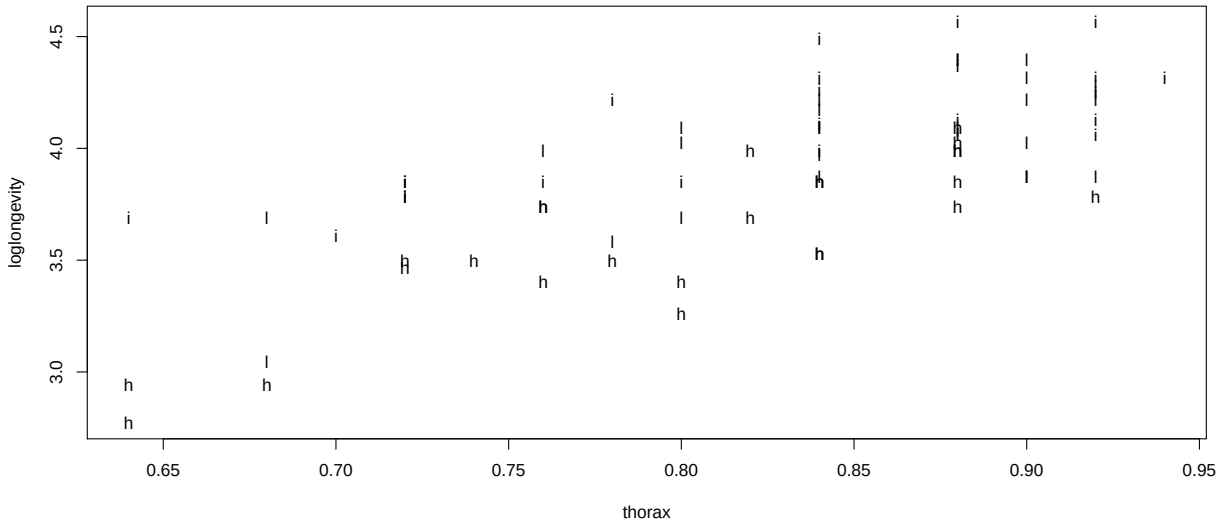
Exercise 1

To investigate the effect of sexual activity on longevity of fruit flies, 75 male fruit flies were divided randomly in three groups of 25. The fruit flies in the first group were kept solitary, those in the second were kept together with one virgin female fruit fly per day, and those in the third group were kept together with eight virgin female fruit flies a day. In the data-file fruitflies.txt the three groups are labelled isolated, low and high. The number of days until death (longevity) was measured for all flies. Later, it was decided to measure also the length of their thorax. Add a column loglongevity to the data-frame, containing the logarithm of the number of days until death. Use this as the response variable in the following.

a) Make an informative plot of the data. Investigate whether sexual activity influences longevity by performing a statistical test, without taking the thorax length into account. What are the estimated longevities for the three conditions? Comment.

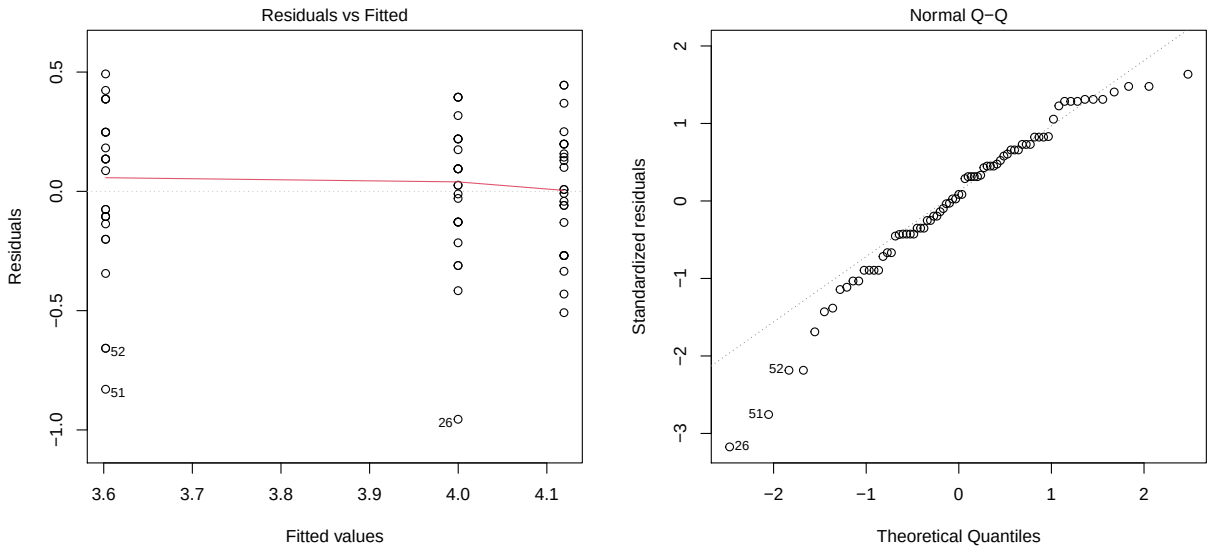
```
data_flies <- read.table(file="data/fruitflies.txt", header=TRUE)
data_flies$activity <- as.factor(data_flies$activity)
# add loglongevity
data_flies <- data_flies %>% mutate(loglongevity = log(longevity))

plot(loglongevity~thorax,pch=as.character(activity), data=data_flies)
```



In the scatter plot there seems to be a positive relationship between thorax and loglongevity, however any obvious influence of the sexual activity level can not be observed.

```r
# perform test to see if sexual activity has an effect on longevity
model <- lm(loglongevity~activity, data = data_flies) # prepare model
par(mfrow=c(1,2)); plot(model, 1); plot(model, 2) # investigate normality
```




```r
anova(model); summary(model)$coefficients
```

```
## Analysis of Variance Table
##
## Response: loglongevity
##           Df Sum Sq Mean Sq F value  Pr(>F)
## activity   2   3.67   1.833    19.4 1.8e-07 ***
## Residuals 72   6.80   0.094
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.602     0.0614   58.62 1.65e-62
## activityisolated    0.517     0.0869    5.95 8.82e-08
## activitylow         0.398     0.0869    4.58 1.93e-05
```

One-way ANOVA was performed to investigate whether sexual activity has an effect on loglongevity. From the results we can see that the p-value < 0.05 meaning that sexual activity level significantly influences loglongevity. From the summary table we can see that all estimates are significantly different from 0: for high sexual activity the estimate is $\exp(3.602)$, for isolated it is $\exp(3.602 + 0.517) = \exp(4.119)$ and for low it is $\exp(3.602 + 0.398) = \exp(4)$.

Test diagnostics: no relationship can be observed in the residuals vs fitted plot. QQ-plot seems to follow a straigt line, however there are some outliers at the extremes.

b) Investigate whether sexual activity influences longevity by performing a statistical test, now including thorax length as an explanatory variable into the analysis. Does sexual activity increase or decrease longevity? What are the estimated longevities for the three groups, for a fly with average thorax length?

```r
# perform additive ANCOVA analysis
model_1 <- lm(loglongevity~thorax+activity, data = data_flies) # prepare model
anova(model_1); table <- summary(model_1)$coefficients; table
```

```
## Analysis of Variance Table
##
## Response: loglongevity
##           Df Sum Sq Mean Sq F value Pr(>F)
## thorax     1   5.43    5.43   132.2 <2e-16 ***
## activity   2   2.11    1.06    25.7  4e-09 ***
## Residuals 71   2.92    0.04
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.219     0.2486    4.90 5.79e-06
## thorax               2.979     0.3067    9.71 1.14e-14
## activityisolated     0.410     0.0584    7.02 1.07e-09
## activitylow          0.286     0.0585    4.88 6.18e-06
```

```r
# extract model's parameter
intercept <- table[,1][1]; beta <- table[,1][2]
alpha_high <- 0; alpha_low <- table[,1][4]
alpha_isolated <- table[,1][3]
# calculate mean thorax
mean_thorax <- mean(data_flies$thorax)
# calculate estimates
estimate_high <- exp(1)**(intercept + alpha_high + beta * mean_thorax)
estimate_low <- exp(1)**(intercept + alpha_low + beta * mean_thorax)
estimate_isolated <- exp(1)**(intercept + alpha_isolated + beta * mean_thorax)
estimates <- c(estimate_isolated, estimate_low, estimate_high)
activity_levels <- unique(as.character(data_flies$activity))
knitr::kable(data.frame(Activity = activity_levels,
                        `Longevity estimate` = estimates),
             caption = "Longevity estimates for average thorax fruit fly")
```

Table 1: Longevity estimates for average thorax fruit fly

Activity	Longevity.estimate
isolated	59.5
low	52.5
high	39.5

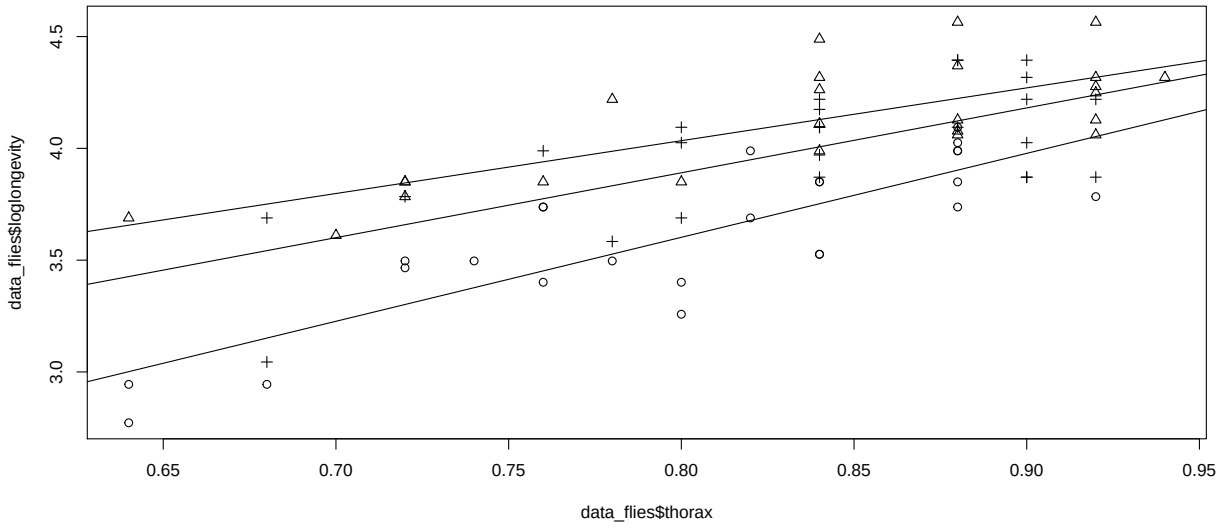
From the ANCOVA analysis results above, we can see that sexual activity has a significant effect (p-values < 0.05) on the loglongevity. From the estimates in the summary table, we can see that sexual activity decreases longevity of the fruit flies - the estimates from isolated and low sexual activity levels are positive with isolated having the highest estimate. Longevity estimates for average thorax fruit fly were estimated by calculating average thorax length (X) and extracting intercept (μ), β and α parameters from the model summary table - the values were plugged into the formula below:

$$Y \approx \mu + \alpha + \beta X$$

The estimates for longevity can be seen in Table 1.

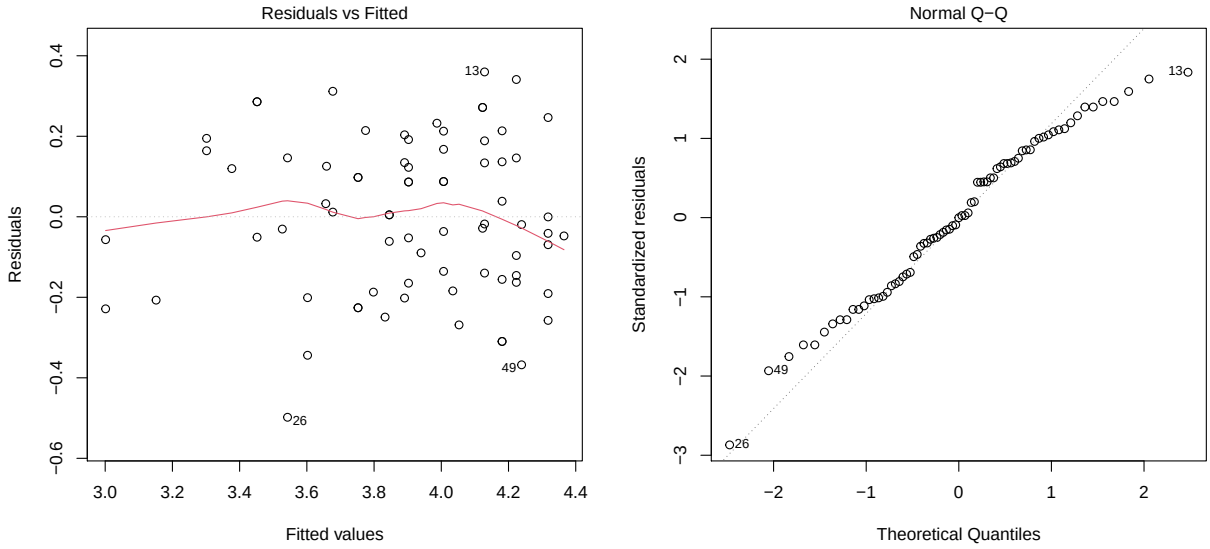
c) How does thorax length influence longevity? Investigate graphically and by using an appropriate test whether this dependence is similar under all three conditions of sexual activity.

```
plot(data_flies$loglongevity~data_flies$thorax,pch=unclass(data_flies$activity))
for (i in activity_levels) {
  abline(lm(loglongevity~thorax
            ,data=data_flies[data_flies$activity==i,]))}
```



From the plot above we can see a positive relationship between thorax and longevity.

```
# perform ANCOVA with interaction analysis
model_interaction <- lm(loglongevity~activity*thorax, data = data_flies) # prepare model
par(mfrow=c(1,2)); plot(model_interaction , 1); plot(model_interaction , 2) # investigate normality
```




```
anova(model_interaction)
```

```
## Analysis of Variance Table
##
## Response: loglongevity
##                  Df Sum Sq Mean Sq F value  Pr(>F)
## activity          2   3.67    1.83   45.77 2.2e-13 ***
## thorax            1   3.88    3.88   96.83 9.0e-15 ***
## activity:thorax   2   0.15    0.08    1.93    0.15
## Residuals        69   2.76    0.04
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

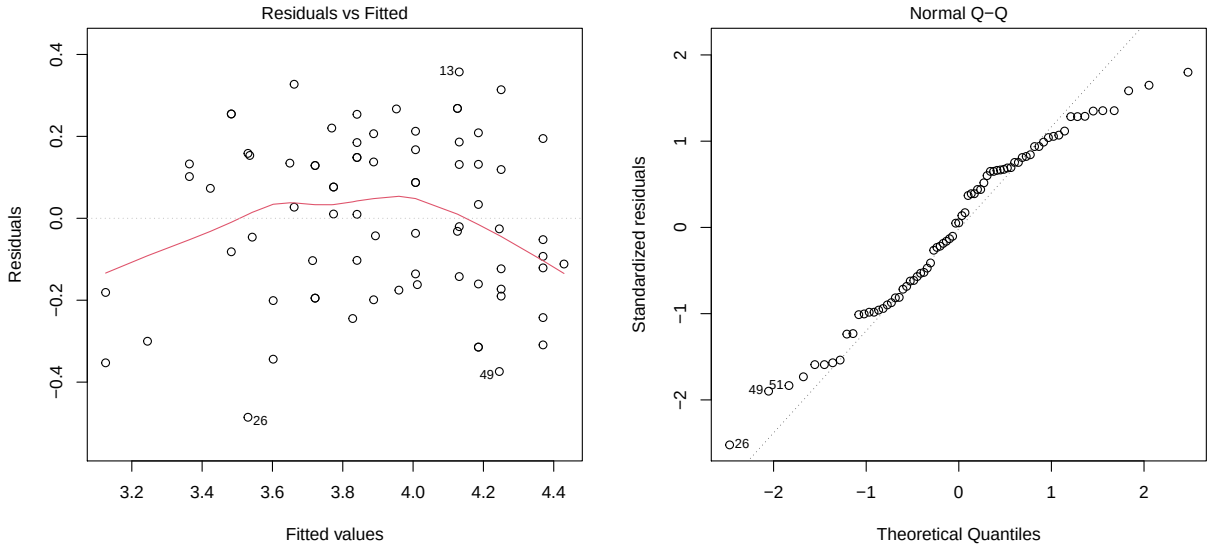
To investigate if this relationship is similar between different sexual activity levels, we need to estimate whether the β parameter (slope) is different between sexual activity levels. From the plot it is not obvious if this is the case - the slopes look very similar. To concretely say if the slopes are the same we performed an ANCOVA analysis with interaction. From the results we can see that the interaction factor is insignificant (p-value > 0.05) and can be ignored. Therefore, the slope parameter can be regarded as the same between different sexual activity levels.

d) Which of the two analyses, without or with thorax length, do you prefer? Is one of the analyses wrong?

Analysis with thorax length is preferable. From the results in *b*) we can see that thorax has a significant influence on the outcome of longevity, therefore it can not be ignored. By performing one-way ANOVA we ignore this influence as it gets absorbed into the single factor of sexual activity level.

e) Verify normality and heteroscedasticity by making a normal QQ-plot of the residuals, and a residuals versus fitted plot, for the analysis that includes thorax length.

```
par(mfrow=c(1,2)); plot(model_1 , 1); plot(model_1 , 2) # investigate normality
```



There does not seem to be any obvious relationship in the Residuals vs Fitted plot. The qq-plot does not follow a straight line well, its shape resembles a letter S, therefore the normality here is questionable.

f) Perform the ancova analysis with the number of days as the response, rather than its logarithm. Verify normality and homoscedasticity of the residuals of this analysis. Was it wise to use the logarithm as response?

```
# perform additive ANCOVA analysis
model <- lm(longevity~thorax+activity, data = data_flies) # prepare model
anova(model); table <- summary(model)$coefficients; table
```

```
## Analysis of Variance Table
##
## Response: longevity
##           Df Sum Sq Mean Sq F value  Pr(>F)
## thorax     1  10959   10959     101 2.6e-15 ***
## activity   2   4967    2483      23 2.0e-08 ***
## Residuals 71   7673     108
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -67.4      12.75   -5.28 1.33e-06
## thorax              132.6      15.72    8.43 2.62e-12
## activityisolated     20.1       2.99    6.70 4.13e-09
## activitylow          13.1       3.00    4.35 4.43e-05
```

```
# extract model's parameter
intercept <- table[,1][1]; beta <- table[,1][2]
alpha_high <- 0; alpha_low <- table[,1][4]
alpha_isolated <- table[,1][3]
# calculate mean thorax
mean_thorax <- mean(data_flies$thorax)
# calculate estimates
```

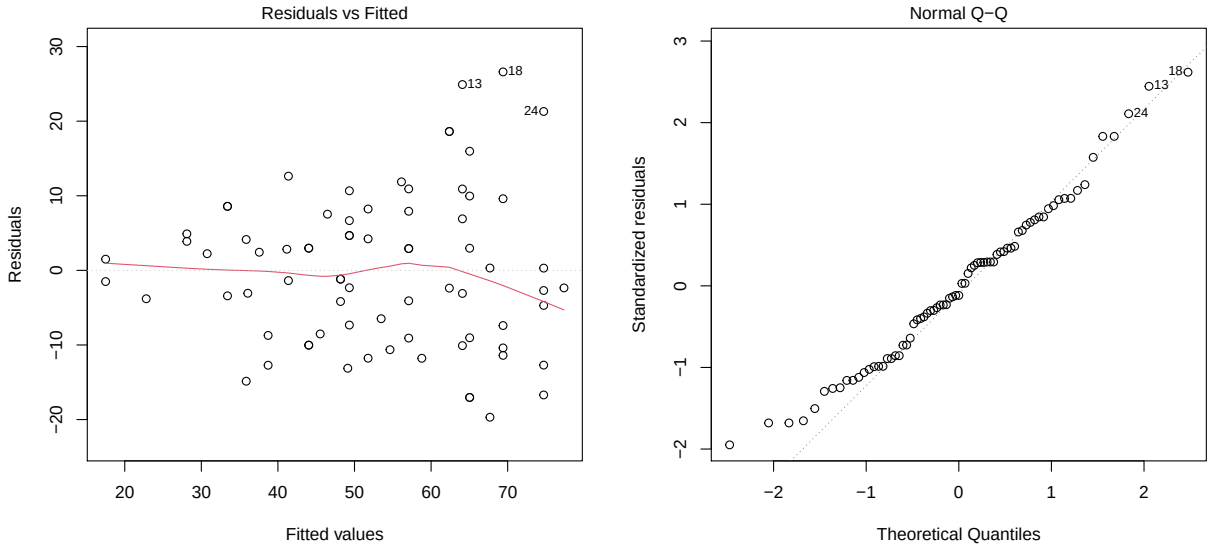
```
estimate_high <- intercept + alpha_high + beta * mean_thorax
estimate_low <- intercept + alpha_low + beta * mean_thorax
estimate_isolated <- intercept + alpha_isolated + beta * mean_thorax
estimates <- c(estimate_isolated, estimate_low, estimate_high)
activity_levels <- unique(as.character(data_flies$activity))
knitr::kable(data.frame(Activity = activity_levels,
                        `Longevity estimate` = estimates),
             caption = "Longevity estimates for average thorax fruit fly")
```

Table 2: Longevity estimates for average thorax fruit fly

Activity	Longevity.estimate
isolated	62
low	55
high	42

The model above brings us to the same conclusion as the model used in b): there is significant influence of sexual activity level on longevity (p-value < 0.05). However, the longevity estimates for average thorax fruit fly for the different levels of sexual activity are slightly different (Table 2).

```
par(mfrow=c(1,2)); plot(model , 1); plot(model , 2) # investigate normality
```



QQ-plot seems to be following a straight line better than the additive model with loglongevity. No obvious relationship can be observed in the Residuals vs Fitted plot and there seems to be less movement here than in the model with loglongevity. Based on the diagnostics, this model with regular longevity better follows the required assumptions. Therefore, it was not wise to logaritmicaly transform longevity.

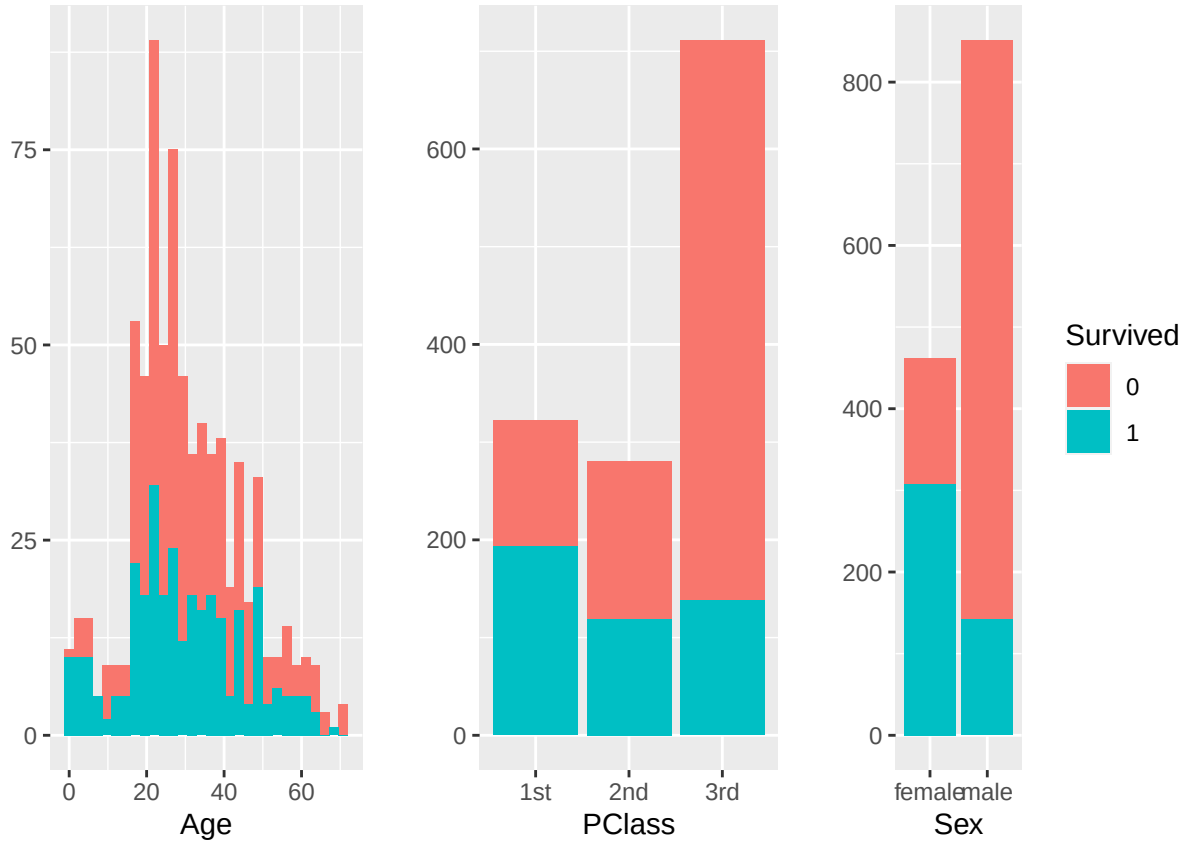
Exercise 2

On April 15, 1912, British passenger liner Titanic sank after colliding with an iceberg. There were not enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew. The data file titanic.txt gives the survival status of passengers on the Titanic, together with their names, age, sex and passenger class. (About half of the ages for the 3rd class passengers are missing, although many of these could be filled in from the original source.) The columns: Name { name of passenger; PClass - passenger class (1st, 2nd or 3rd), Age - age in years, Sex - male or female, Survived - survival status (1=Yes or 0=No).

a) Study the data and give a few (>1) summaries (graphics or tables).

```
par(mfrow=c(1,3))
titanic <- read.table(file="data/titanic.txt", header=TRUE)
titanic$Survived = as.factor(titanic$Survived)
plot1 <- qplot(x = Age, fill = Survived, data = titanic, show.legend = FALSE)
plot2 <- qplot(x = PClass, fill = Survived, data = titanic, show.legend = FALSE)
plot3 <- qplot(x = Sex,fill=Survived, data = titanic)
grid.arrange(plot1, plot2, plot3, ncol=3)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
titanic <- read.table(file="data/titanic.txt", header=TRUE)
class_sex =  xtabs(~PClass+ Sex, data=titanic);knitr::kable(class_sex)
```

	female	male
1st	143	179
2nd	107	173
3rd	212	499

```
tot = xtabs(Survived ~ PClass + Sex, data=titanic);knitr::kable(tot)
```

	female	male
1st	134	59
2nd	94	25
3rd	80	58

```
knitr::kable(round(tot/class_sex, 2))
```

	female	male
1st	0.94	0.33
2nd	0.88	0.14
3rd	0.38	0.12

The histogram shows the ages of the passenger and whether or not they survived the Titanic accident or not. The two barplots show the survivors and perished people with respect to sex and class. We can see already here that the fraction of survivors in the first class is bigger than the 2nd and 3rd class and the fraction of survivors in the 3rd class is the least. The barplot of the survivors with respect to the sex show that the fraction of male survivors is much less than female survivors. These fractions can also be seen in the tables which shows, categorized on class and sex, the the total amount of people on board of the titanic, the survivors and the fractions.

b) Fit a logistic regression model (no interactions yet) to investigate the association between the survival status and the predictors PClass, Age and Sex. Interpret the results in terms of odds, comment.

```
titanic$PClass <- as.factor(titanic$PClass)
titanic$Sex <- as.factor(titanic$Sex)
logistic <- glm(Survived~PClass+Age+Sex, data = titanic, family = binomial)
summary(logistic)$coefficients
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.7597    0.39757    9.46 3.18e-21
## PClass2nd    -1.2920    0.26008   -4.97 6.78e-07
## PClass3rd    -2.5214    0.27666   -9.11 7.95e-20
## Age          -0.0392    0.00762   -5.14 2.69e-07
## Sexmale      -2.6314    0.20151  -13.06 5.68e-39
```

```
drop1(logistic,test="Chisq")
```

```
## Single term deletions
##
## Model:
```

```
## Survived ~ PClass + Age + Sex
##        Df Deviance AIC   LRT Pr(>Chi)
## <none>         695 705
## PClass  2      796 802 100.4  < 2e-16 ***
## Age     1      724 732  28.5  9.6e-08 ***
## Sex     1      910 918 214.8  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
odds_male_example =  exp(3.7597 + 1 * -1.2920 + 0 * -2.5214 + 1 * -2.6314 + 25 * -0.0392)
odds_male_example
```

```
## [1] 0.319
```

```
odds_female_example =  exp(3.7597 + 0 * -1.2920 + 0 * -2.5214 + 0 * -2.6314 + 25 * -0.0392)
odds_female_example
```

```
## [1] 16.1
```

The odds can be defined as the probability of success devided by the probability of failure. The summary show that the odds of surviving is exp(3.7597 + PClass2nd * -1.2920 + PClass3rd * -2.5214 + Sexmale * -2.6314 + age * -0.0392). This means that the odds of survival while being a female in the first class is exp(3.7597 + age * -0.0392). This shows that the odds of survival for a man who is 25 years old in the 2nd class is 0.319. While a woman who is in the first class and has an age of 25 has the odds of survival of 16.1. The drop1 table also shows that all estimators are significant. The model contains some uncertainty due to to the fact that about half of the ages for the 3rd class passengers are missing. These data points will not be taken into account while the model was created. **c)** Investigate for interaction of predictor Age with factors PClass and Sex. From this and b), choose (and justify) a resulting model. For this model, report the estimate for the probability of survival for each combination of levels of the factors PClass and Sex for a person of age 53.

```
# interaction of sex and age
logistic_interaction_sex <- glm(Survived~Age*Sex, data = titanic, family = binomial)
summary(logistic_interaction_sex)$coefficients
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.3011     0.2990    1.01 3.14e-01
## Age           0.0294     0.0101    2.91 3.58e-03
## Sexmale      -0.5999     0.4080   -1.47 1.42e-01
## Age:Sexmale  -0.0657     0.0137   -4.80 1.57e-06
```

```
# interaction of PClass and age
logistic_interaction_pclass <- glm(Survived~Age*PClass, data = titanic, family = binomial)
summary(logistic_interaction_pclass)$coefficients
```

```
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.92298    0.43625   4.408 1.04e-05
## Age           -0.03584    0.00996  -3.600 3.18e-04
## PClass2nd     -0.74428    0.57155  -1.302 1.93e-01
## PClass3rd     -2.29007    0.54057  -4.236 2.27e-05
## Age:PClass2nd -0.01321    0.01587  -0.832 4.05e-01
## Age:PClass3rd  0.00464    0.01594   0.291 7.71e-01
```

Two logistic regressions with interaction were performed to investigate the interaction between Age-Sex and Age-PClass. Significant interaction was found between Age and Sex, therefore it will be added to the model. No significant interaction between Age and PClass was identified.

```
# add the interaction term
logistic_interaction_1 <- glm(Survived~PClass+Age+Sex+Age:Sex, data = titanic,
                              family = binomial) # remove Age

logistic_interaction_2 <- glm(Survived~PClass+Sex+Age:Sex, data = titanic,
                              family = binomial) # remove Sex

logistic_interaction_final <- glm(Survived~PClass+Age:Sex, data = titanic, family = binomial)
summary(logistic_interaction_final)$coefficients
```

```
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)      2.5040    0.37777    6.63 3.40e-11
## PClass2nd       -1.5810    0.28802   -5.49 4.04e-08
## PClass3rd       -2.6661    0.29235   -9.12 7.54e-20
## Age:Sexfemale    0.0108    0.00894    1.21 2.27e-01
## Age:Sexmale     -0.0802    0.00917   -8.74 2.24e-18
```

After adding the interaction term into the model (PClass+Age+Sex+Age:Sex) we observed that Sex and Age were no longer significant variables, therefore they were removed from the model. The resulting model - PClass+Age:Sex - is the final model we chose. This is the preferred model over the one used in *b*) as that model ignores significant interaction of the variables.

```
# predict for 53 years and all PClass, Sex
classes <- as.character(unique(titanic$PClass))
sexes <- as.character(unique(titanic$Sex))
age <- 53
new_data <- expand.grid(PClass = classes, Sex = sexes, Age = age)
results <- predict(logistic_interaction_final, new_data, type="response")
final <- new_data %>% bind_cols(Survival = results)
knitr::kable(final)
```

PClass	Sex	Age	Survival
1st	female	53	0.956
2nd	female	53	0.817
3rd	female	53	0.601
1st	male	53	0.148
2nd	male	53	0.035
3rd	male	53	0.012

The survival predictions for a 53 year-old passenger of all PClass and Sex can be observed above.

d) Propose a method to predict the survival status and a quality measure for your prediction and describe how you would implement that method (you do not need to implement it).

Split data into training and testing data. Train the logistic regression model on the training data and then use this model to predict the outcomes of the unseen, testing data. Set a threshold for the probability to be converted to success/fail - percentage correct could be the quality measure for the model.

e) Another approach would be to apply a contingency table test to investigate whether factor passenger class (and gender) has an effect on the survival status. Implement the relevant test(s).

```
# Makes a contingency table and chi-squared test for survived with class and gender
tab1 = table(titanic$Survived,titanic$Sex)
fisher.test(tab1)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tab1
## p-value <2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.0762 0.1316
## sample estimates:
## odds ratio
##        0.1
```

```
total_s= rowSums(rbind(tab1,NaN));total_g = colSums(tab1);
tab1 = rbind(tab1,total_g);tab1 = cbind(tab1,total_s)
knitr::kable(tab1)
```

	female	male	total_s
0	154	709	863
1	308	142	450
total_g	462	851	NaN

```
tab2 = table(titanic$Survived,titanic$PClass)
chisq.test(tab2)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab2
## X-squared = 172, df = 2, p-value <2e-16
```

```
total_c= rowSums(tab2);total_g = colSums(tab2);
tab2 = rbind(tab2,total_g);tab2 = cbind(tab2,total_s)
knitr::kable(tab2)
```

	1st	2nd	3rd	total_s
0	129	161	573	863
1	193	119	138	450
total_g	322	280	711	NaN

The Fischer test is used for the contingency tables when the effect of sex on survival is investigated because. This could be done because it us a 2x2 contingency table. For the influence of the class on survival a Chi-

square test is conducted because of the 2x3 table, this test reliable because more than 80% of the values have a value of more than 5. The tests tests above we show that both factors have a significant effect on the survival outcome.

f) Is the second approach in e) wrong? Why or why not? Name both an advantage and a disadvantage of the two approaches, relative to each other.

The second approach is not necessary wrong as we want to know whether or not the Survived variable is independent of the variables gender and class. The approach in e (the contingency table with the Fisher and chi-squared test) is for that question suitable.

Whether you want to use the contingency table with the chi-squared test depends generally on which question we want to answer. A advantage of the Chi-squared is that the test can be used when we want to know if the reason that a person survived is independent of their sex or class, as mentioned above. While the logistic regression has the advantage that it could compute the probability that a person with a certain, age, class and gender survives or not.

Exercise 3

To study the influence of different political and geographical variables on the number of military coups, these data are collected for several Sub Saharan African countries in the file africa.txt. The meaning of the different variables:

miltcoup - number of successful military coups from independence to 1989; oligarchy - number years country ruled by military oligarchy from independence to 1989; pollib - political liberalization (0 = no civil rights for political expression, 1 = limited civil rights for expression but right to form political parties, 2 = full civil rights); parties - number of legal political parties in 1993; pctvote - percent voting in last election; popn - population in millions in 1989; size - area in 1000 square km; numelec - total number of legislative and presidential elections; numregim - number of regime types.

a) Perform Poisson regression on the full data set africa, taking miltcoup as response variable, Comment on your findings.

```
africa <- read.table(file="data/africa.txt", header=TRUE)
africa$pollib <- as.factor(africa$pollib)
poisson_model <- glm(miltcoup~oligarchy+pollib+parties+pctvote+popn+size+numelec+numregim,
                     family=poisson,data=africa)
summary(poisson_model)
```

```
##
## Call:
## glm(formula = miltcoup ~ oligarchy + pollib + parties + pctvote +
##     popn + size + numelec + numregim, family = poisson, data = africa)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.508  -0.953  -0.310   0.486   1.646
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.233427   0.997611   -0.23   0.8150
## oligarchy    0.072566   0.035346    2.05   0.0401 *
## pollib1     -1.103244   0.655811   -1.68   0.0925 .
## pollib2     -1.690306   0.676650   -2.50   0.0125 *
```

```
## parties      0.031221   0.011166    2.80   0.0052 **
## pctvote      0.015441   0.010103    1.53   0.1264
## popn         0.010959   0.007149    1.53   0.1253
## size        -0.000265   0.000269   -0.99   0.3244
## numelec     -0.029619   0.069625   -0.43   0.6705
## numregim     0.210943   0.233933    0.90   0.3672
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 65.945  on 35  degrees of freedom
## Residual deviance: 28.249  on 26  degrees of freedom
## AIC: 113.1
##
## Number of Fisher Scoring iterations: 5
```

When inspecting the above summary of the poisson regression for the whole dataset with miltcoup as response variable, we see that only three variables show significant influence. Namely the variables oligarchy, pollib and parties show a p-value < 0.05. For the pollib factor variable we also see that only pollib2 level (full civil rights) has a significant influence. Furthermore, the number of legislative and presidential elections (numelec) and the number of regime types seem to have the least influence on the number of coups.

b) Use the step down approach (using output of the function summary) to reduce the number of explanatory variables. Compare the resulting model with your findings in a).

```
# For every summary the variable with the highest P-value was excluded
summary(glm(miltcoup~oligarchy+pollib+parties+pctvote+popn+size+
              numelec+numregim,family=poisson,data=africa)) # remove numelec

summary(glm(miltcoup~oligarchy+pollib+parties+pctvote+popn+size+
              numregim,family=poisson,data=africa)) # remove numregim

summary(glm(miltcoup~oligarchy+pollib+parties+pctvote+popn+size,
            family=poisson,data=africa)) # remove size

summary(glm(miltcoup~oligarchy+pollib+parties+pctvote+popn,
            family=poisson,data=africa)) # remove popn

summary(glm(miltcoup~oligarchy+pollib+parties+pctvote,
            family=poisson,data=africa)) # remove pctvote
```

```
final_poisson <- glm(miltcoup~oligarchy+pollib+parties,
                     family=poisson,data=africa) # final
summary(final_poisson)
```

```
##
## Call:
## glm(formula = miltcoup ~ oligarchy + pollib + parties, family = poisson,
##     data = africa)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
```

```
## -1.361  -1.041  -0.315   0.615   1.754
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.2080     0.4457    0.47    0.641
## oligarchy     0.0915     0.0226    4.05    5e-05 ***
## pollib1      -0.4954     0.4757   -1.04    0.298
## pollib2      -1.1121     0.4595   -2.42    0.016 *
## parties       0.0224     0.0091    2.46    0.014 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 65.945  on 35  degrees of freedom
## Residual deviance: 32.822  on 31  degrees of freedom
## AIC: 107.6
##
## Number of Fisher Scoring iterations: 5
```

In the final_poisson model we are left with the same three variables as were already flagged as significant at question 3a, namely: oligarchy, pollib and parties. And also again only one of the factor levels of pollib (pollib2) remains significant.

c) Predict the number of coups for a hypothetical country for all the three levels of political liberalization and the averages (over all the counties in the data) of all the other (numerical) characteristics. Comment on your findings.

```
# get average values
new_data <- africa %>% mutate_if(is.numeric, mean) %>% select(-pollib, -miltcoup)
new_data <- new_data[1,]
avg_oligarchy <- new_data$oligarchy; avg_parties <- new_data$parties;

# get a list of political liberalization levels
pollib <- sort(as.character(unique(africa$pollib)))
new_data_reduced <- expand.grid(pollib = pollib, oligarchy = avg_oligarchy,
                        parties = avg_parties)

results <- predict(final_poisson, new_data_reduced, type="response")
final <- new_data_reduced %>% bind_cols(Prediction = results)
knitr::kable(final, caption = "Predictions with the model from b)")
```

Table 9: Predictions with the model from b)

pollib	oligarchy	parties	Prediction
0	5.22	17.1	2.908
1	5.22	17.1	1.772
2	5.22	17.1	0.956

Looking at the predictions we see that the amount of expected coups decreases with higher levels of political liberalization. In both cases full civil rights (pollib = 2) predicts the lowest number of coups < 1.