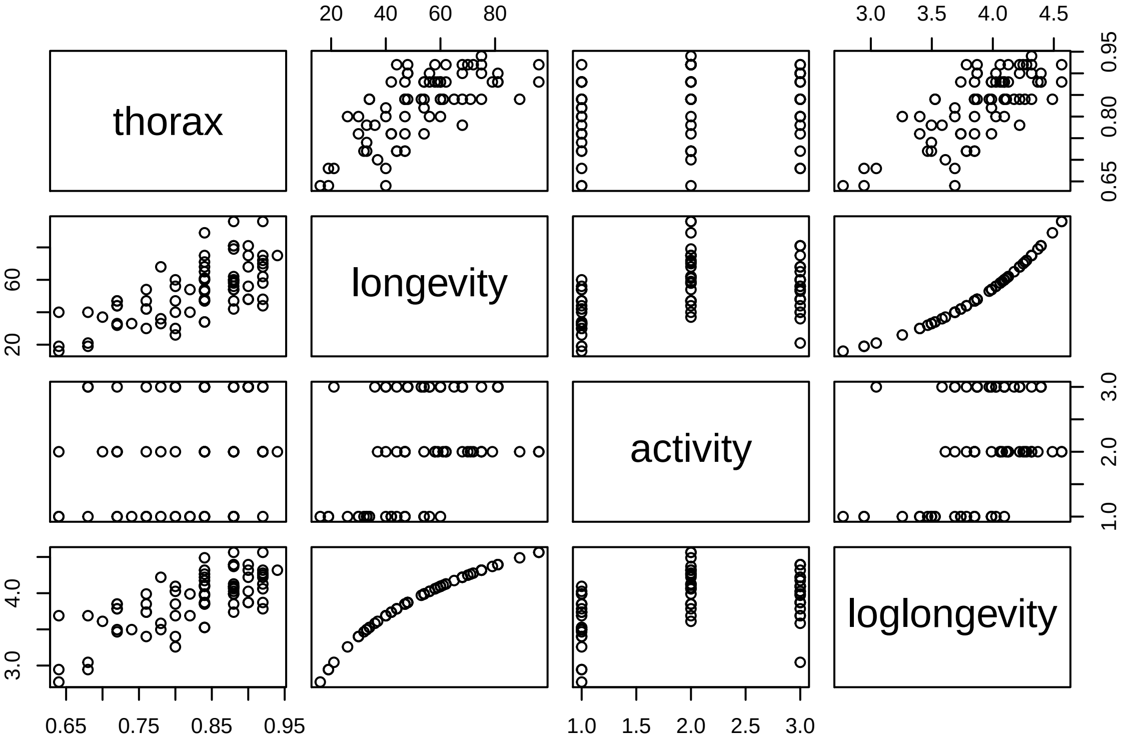
EDDA - Assignment 3 - Group 77

Dante de Lang, Ignas Krikštaponis and Kamiel Gülpen

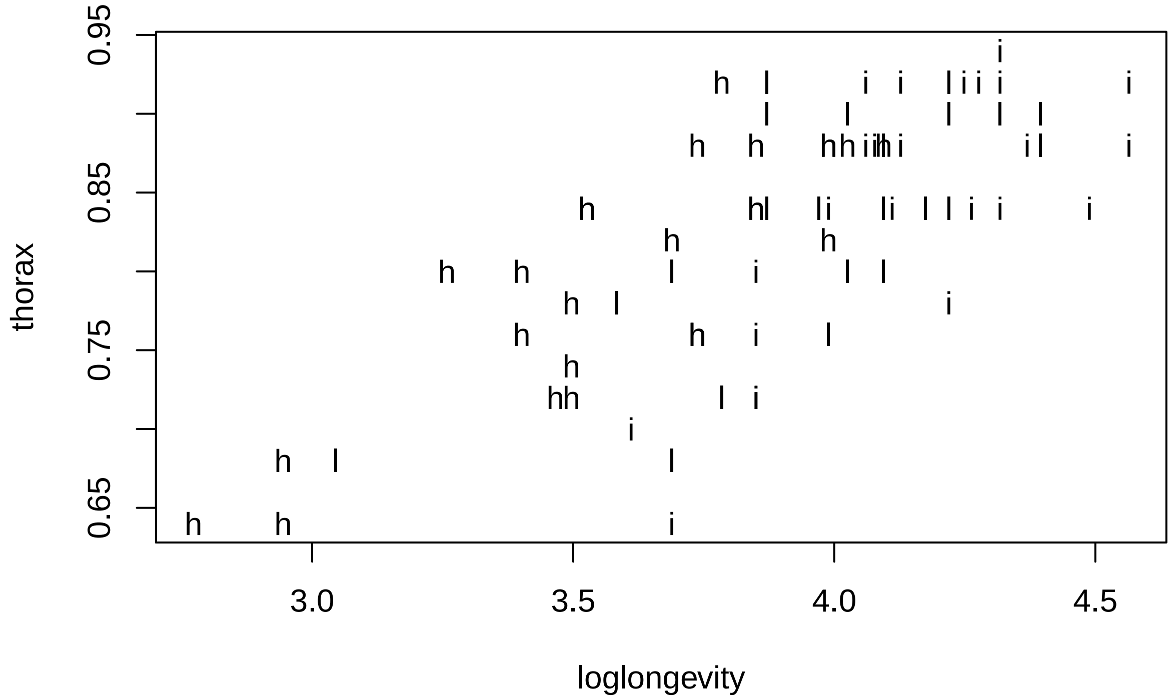
Exercise 1

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.

```r
data_flies <- read.table(file="data/fruitflies.txt", header=TRUE)

# Add log colomn
data_flies$loglongevity <- log(data_flies$longevity)

plot(data_flies)
```



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



```
data_flies$activity <- as.factor(data_flies$activity)
flies_lm_1 <- lm(loglongevity~activity, data=data_flies)

anova(flies_lm_1)
```

```
## 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
```

```
summary(flies_lm_1)
```

```
##
## Call:
## lm(formula = loglongevity ~ activity, data = data_flies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9553 -0.1334  0.0255  0.2089  0.4922
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.6021     0.0614   58.62  < 2e-16 ***
## activityisolated   0.5172     0.0869    5.95  8.8e-08 ***
## activitylow        0.3977     0.0869    4.58  1.9e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.307 on 72 degrees of freedom
## Multiple R-squared:  0.35,   Adjusted R-squared:  0.332
## F-statistic: 19.4 on 2 and 72 DF,  p-value: 1.8e-07
```

Anova gives significant result of influence activity on longevity (1.8e-7). Summary shows estimated loglongevity for different activity levels: 3.6021 for high, 3.6021+0.5172 for isolated and 3.6021+0.3977 for low. What is very peculiar about these results is that the isolated group without female fruit flies survives for the longest time while the group with high activity dies out first.

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 y with average thorax length?

```
data_flies$activity <- as.factor(data_flies$activity)
flies_lm_2 <- lm(loglongevity~thorax+activity, data=data_flies)

anova(flies_lm_2)
```

```
## 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
```

```
summary(flies_lm_2)
```

```
##
## Call:
## lm(formula = loglongevity ~ thorax + activity, data = data_flies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4858 -0.1612  0.0104  0.1510  0.3574
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.2189     0.2486    4.90  5.8e-06 ***
## thorax             2.9790     0.3067    9.71  1.1e-14 ***
## activityisolated   0.4100     0.0584    7.02  1.1e-09 ***
## activitylow        0.2857     0.0585    4.88  6.2e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.203 on 71 degrees of freedom
## Multiple R-squared:  0.721,  Adjusted R-squared:  0.709
## F-statistic: 61.2 on 3 and 71 DF,  p-value: <2e-16
```

```r
drop1(flies_lm_2, test = "F")
```

```
## Single term deletions
##
## Model:
## loglongevity ~ thorax + activity
##          Df Sum of Sq  RSS  AIC F value  Pr(>F)
## <none>                2.92 -236
## thorax    1      3.88 6.80 -174    94.4 1.1e-14 ***
## activity  2      2.11 5.03 -199    25.7 4.0e-09 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

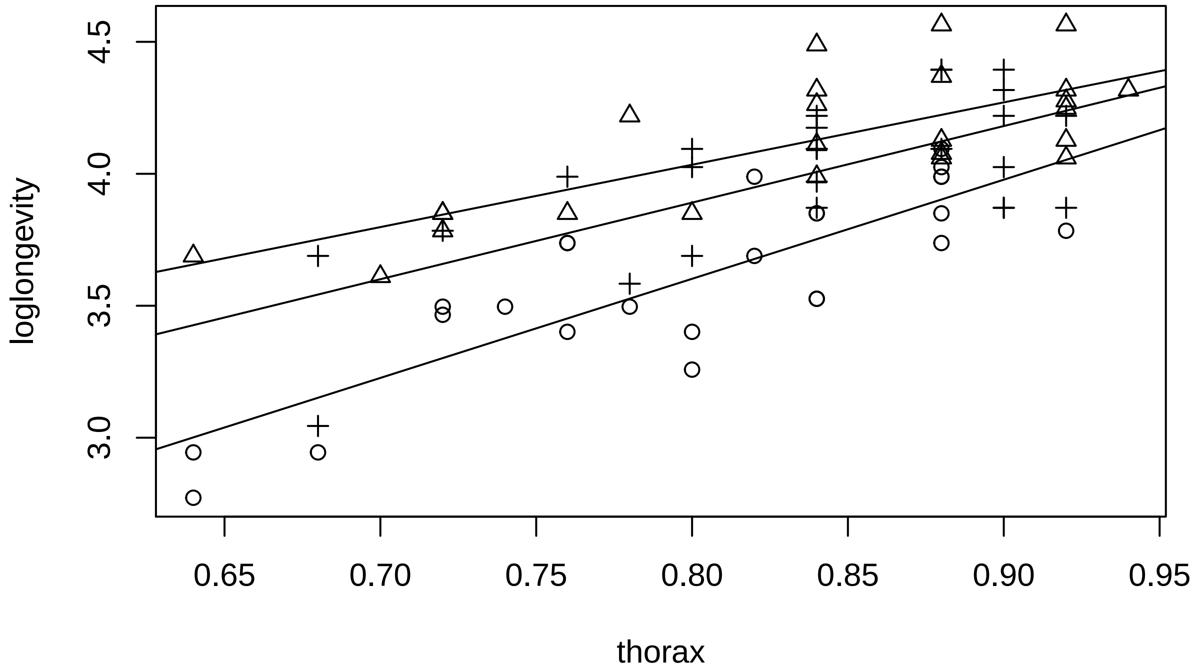
```r
confint(flies_lm_2)
```

```
##                  2.5 % 97.5 %
## (Intercept)      0.723  1.715
## thorax           2.368  3.590
## activityisolated 0.294  0.526
## activitylow      0.169  0.402
```

High sexual activity returns the lowest longevity, therefore we can state it decreases longevity. For the estimated longevities for the three groups we can follow the equation $Y_{in} = \mu + \alpha_i + \beta X_{in} + e_{in}$. This gives us the following values: Isolated = 1.2189+0.4100+2.9790+0.203 = 4.81 Low = 1.2189+0.2857+2.9790+0.203 = 4.69 High = 1.2189+0+2.9790+0.203 = 4.4

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.

```r
plot(loglongevity~thorax,pch = unclass(activity),data=data_flies)
activity_classes <- c("isolated","low","high")
for (class in activity_classes) abline(lm(loglongevity~thorax, data=data_flies[data_flies$activity==cla
```



```
flies_lm_3 <- lm(loglongevity~thorax*activity, data=data_flies)
anova(flies_lm_3)
```

```
## Analysis of Variance Table
##
## Response: loglongevity
##                 Df Sum Sq Mean Sq F value  Pr(>F)
## thorax           1   5.43    5.43  135.62 < 2e-16 ***
## activity         2   2.11    1.06   26.38 3.1e-09 ***
## thorax:activity  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
```

```
summary(flies_lm_3)
```

```
##
## Call:
## lm(formula = loglongevity ~ thorax * activity, data = data_flies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4980 -0.1592 -0.0003  0.1462  0.3598
##
## Coefficients:
```

```
##                           Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                  0.598      0.419    1.43    0.158    
## thorax                       3.755      0.522    7.20  5.8e-10 ***
## activityisolated             1.546      0.584    2.65    0.010 *  
## activitylow                  0.972      0.642    1.51    0.135    
## thorax:activityisolated     -1.393      0.712   -1.96    0.055 .  
## thorax:activitylow          -0.854      0.779   -1.10    0.277    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.2 on 69 degrees of freedom
## Multiple R-squared:  0.736,  Adjusted R-squared:  0.717 
## F-statistic: 38.4 on 5 and 69 DF,  p-value: <2e-16
```

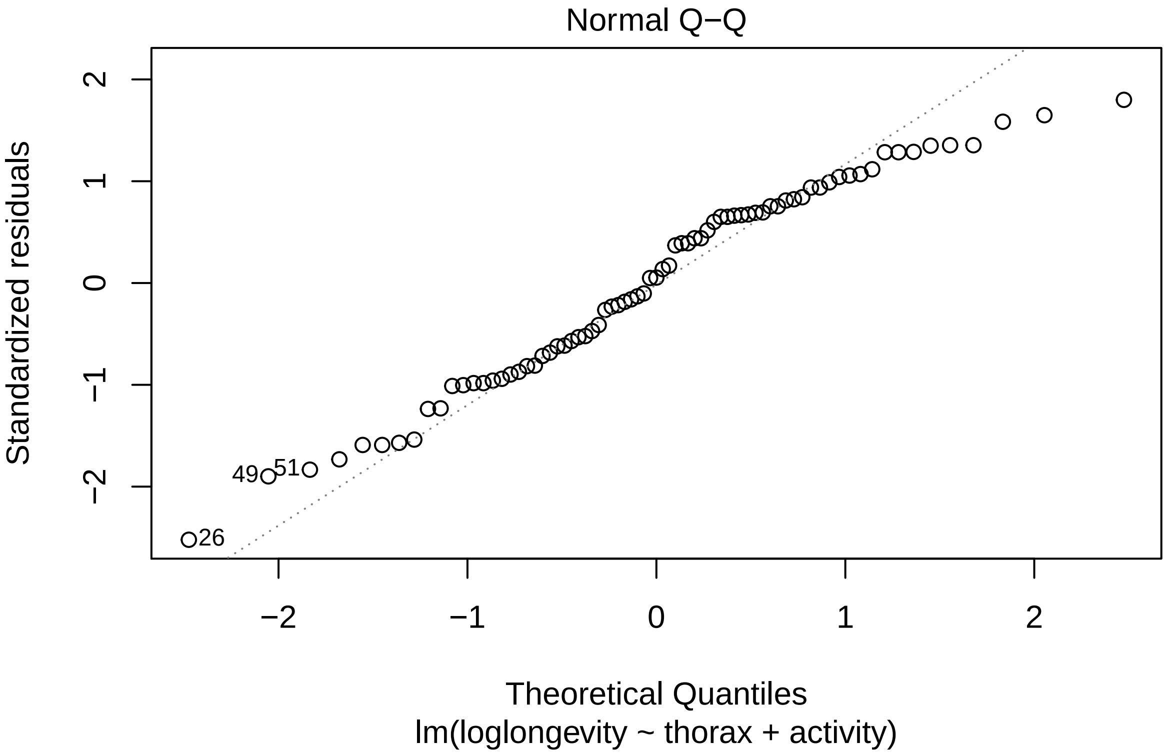
Thorax length seems to have a positive effect on the longevity of the fruit flies, this is visually shown in the plot with the ablines. From the plot this seems to be the case for all three activity groups. When checking for interaction through ANOVA, we see no significant interaction between thorax and activity, therefore our first analysis can be trusted.

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

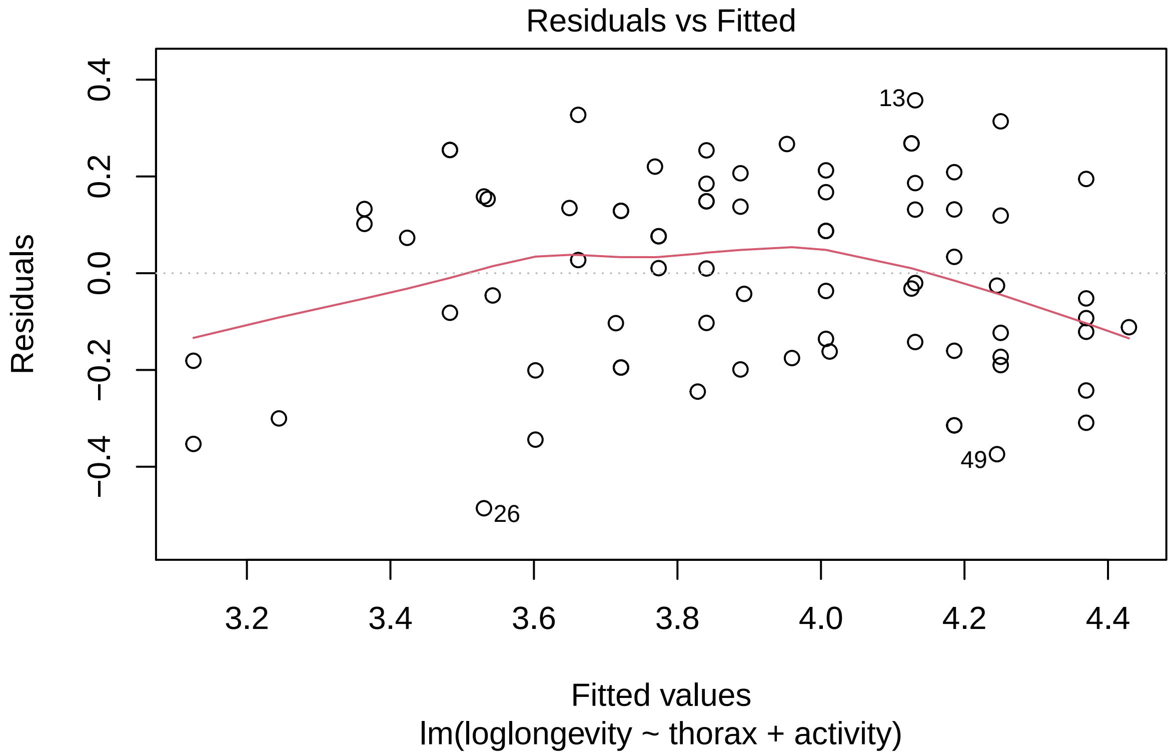
Since there is no interaction between thorax length and activity, I would suggest to use the model without. This will give clearer insights in the influence of activity on longevity. Dunno if last model with * is wrong?!?

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.

```
plot(flies_lm_2, 2)
```



```
plot(flies_lm_2, 1)
```



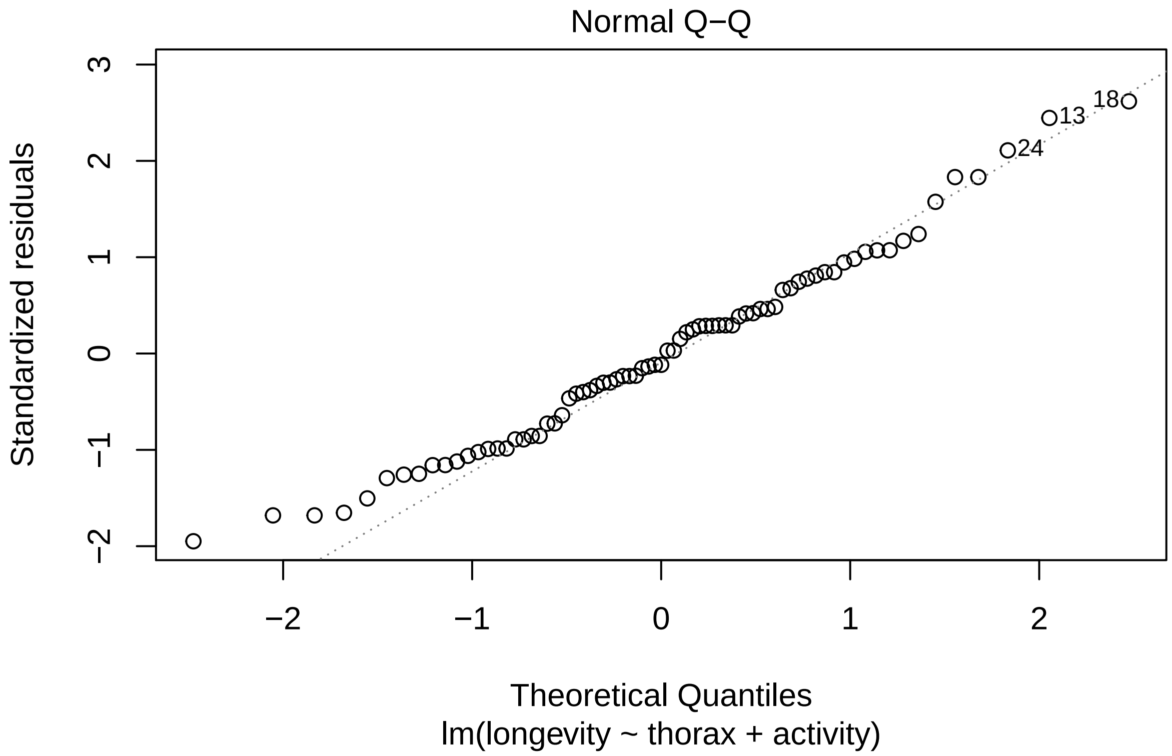
Looking at the QQ-norm plot we can assume normality since the datapoints resemble an almost linear line, only the outer values deviate a bit. For the residuals-fitted plot we can see asymmetry which is an indicator for heteroscedasticity.

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?

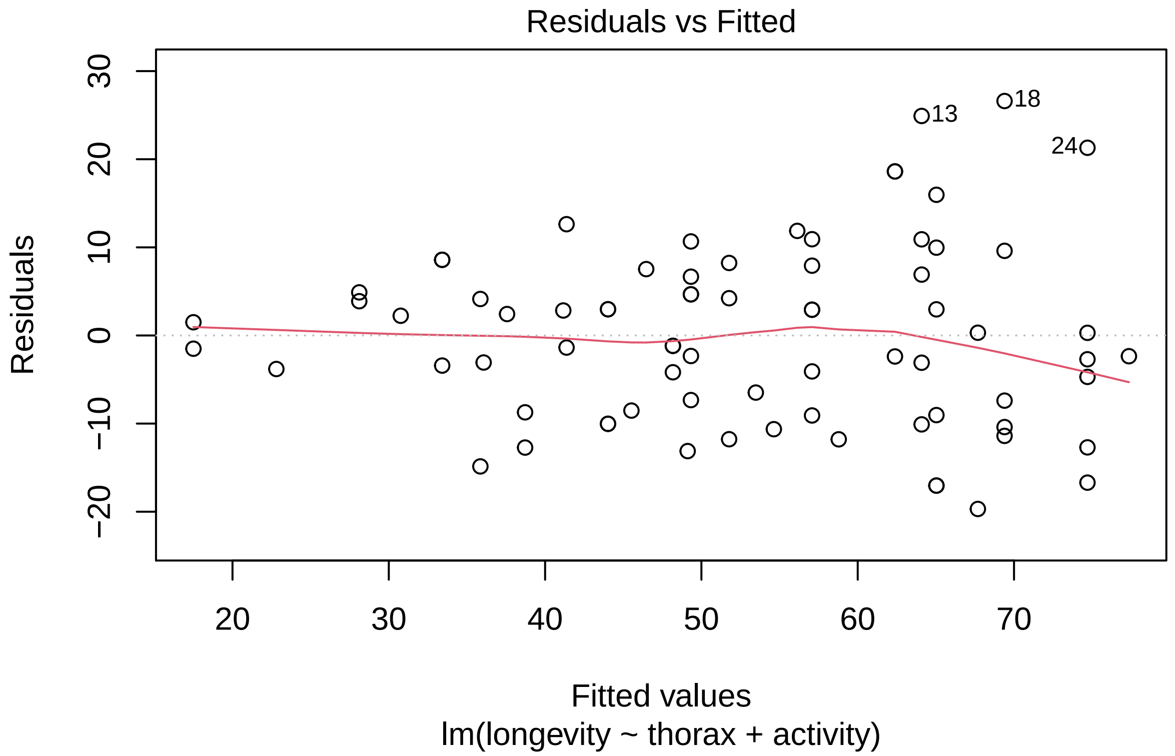
```
flies_lm_4 <- lm(longevity~thorax+activity, data=data_flies)

anova(flies_lm_4)
```

```
## 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
```

```
plot(flies_lm_4, 2)
```



```
plot(flies_lm_4, 1)
```

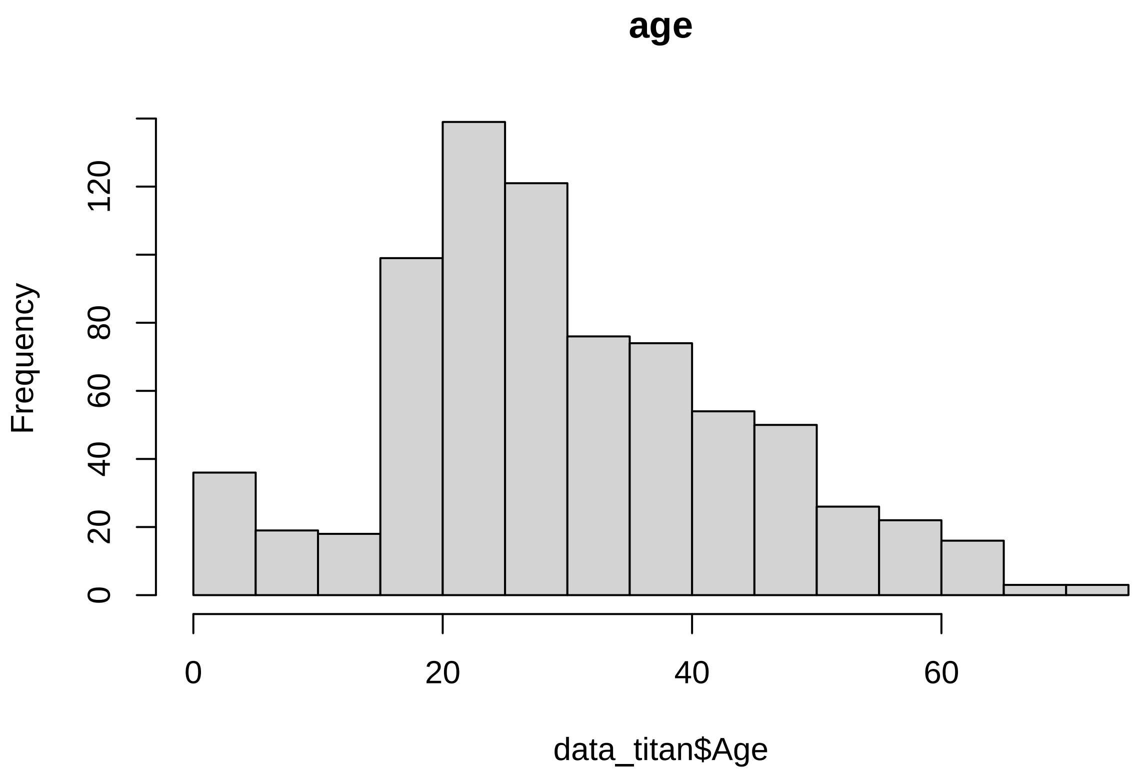


Here the QQ-norm plot again shows normality where only the lowest quantiles deviate form the fitted line. For the residuals-fitted plot we can also see clear symmetry with a fairly straight fitted line, this is then also an indicator for homoscedasticity.

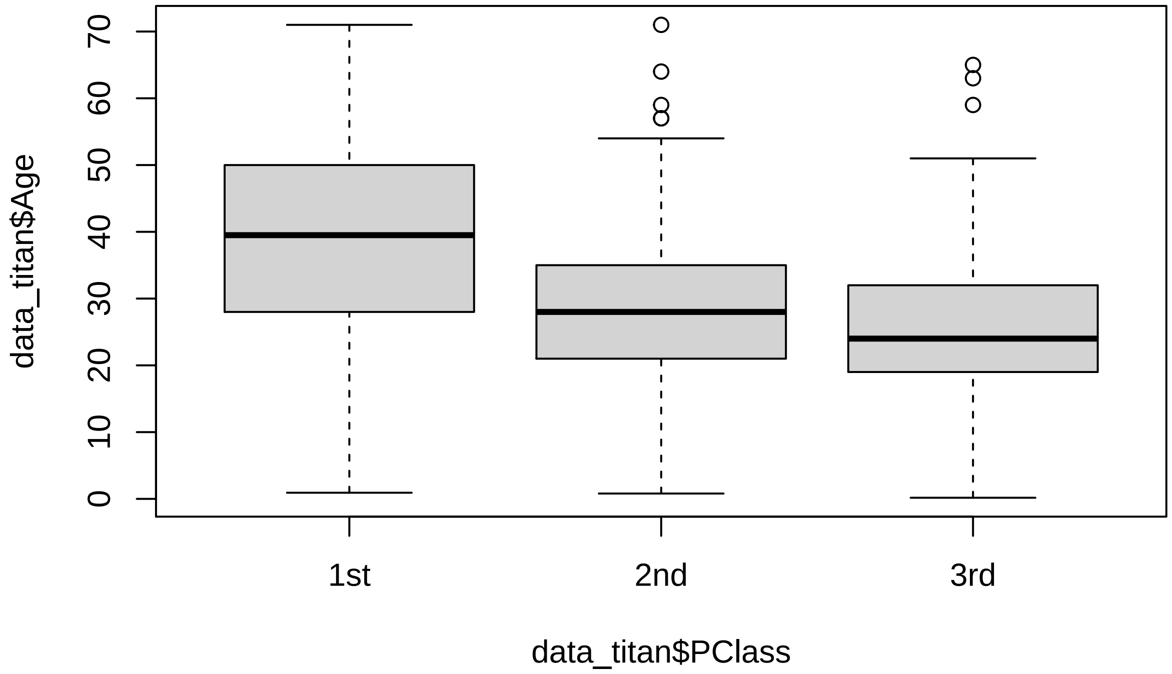
Exercise 2

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

```
data_titan <- read.table(file="data/titanic.txt", header=TRUE)

data_titan$PClass <- as.factor(data_titan$PClass)
data_titan$Sex <- as.factor(data_titan$Sex)


hist(data_titan$Age,main="age")
```



```
boxplot(data_titan$Age~data_titan$PClass)
```



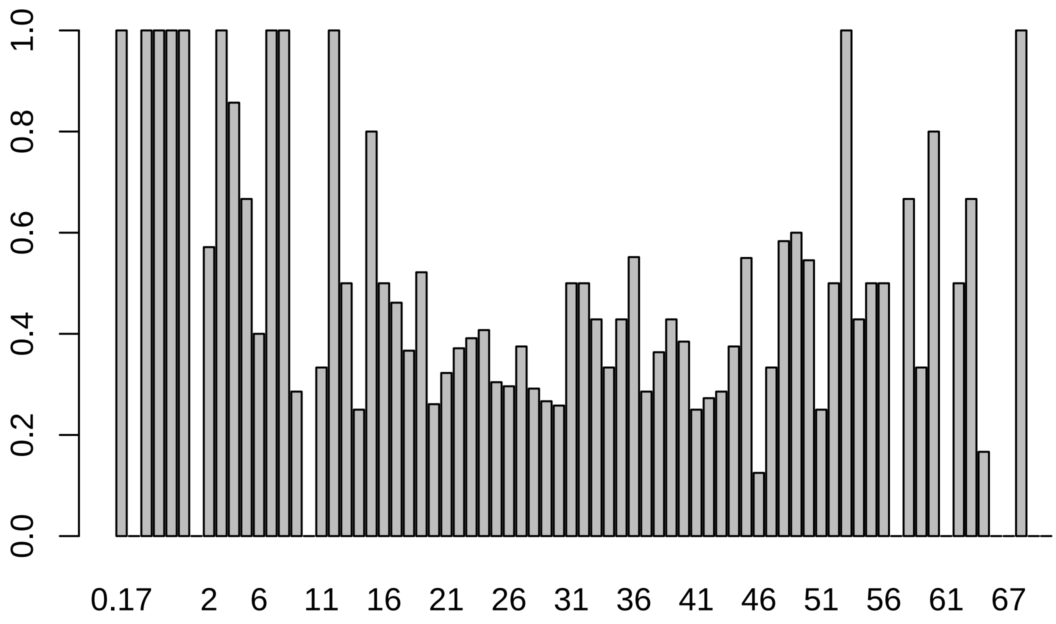
```
titan_table_1 <- xtabs(~Sex+PClass, data=data_titan)
titan_table_1
```

```
##         PClass
## Sex       1st 2nd 3rd
##   female 143 107 212
##   male   179 173 499
```

```
titan_table_1.c <- xtabs(Survived~Sex+PClass, data=data_titan)
round(titan_table_1.c/titan_table_1,2)
```

```
##         PClass
## Sex       1st  2nd  3rd
##   female 0.94 0.88 0.38
##   male   0.33 0.14 0.12
```

```
# Not really nice but perhaps group age partners
titan_table_2 <- xtabs(~Age, data=data_titan)
barplot(xtabs(Survived~Age, data=data_titan)/titan_table_2)
```



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.

```
titan_glm <- glm(Survived~PClass+Age+Sex,data=data_titan,family = binomial)

summary(titan_glm)
```

```
##
## Call:
## glm(formula = Survived ~ PClass + Age + Sex, family = binomial,
##     data = data_titan)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.723  -0.707  -0.392   0.649   2.529
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.75966    0.39757    9.46  < 2e-16 ***
## PClass2nd   -1.29196    0.26008   -4.97  6.8e-07 ***
## PClass3rd   -2.52142    0.27666   -9.11  < 2e-16 ***
## Age         -0.03918    0.00762   -5.14  2.7e-07 ***
## Sexmale     -2.63136    0.20151  -13.06  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1025.57  on 755  degrees of freedom
## Residual deviance:  695.14  on 751  degrees of freedom
##   (557 observations deleted due to missingness)
## AIC: 705.1
##
## Number of Fisher Scoring iterations: 5
```

Looking at the summary of the logistic regression model, we can see that the probability of survival is mostly dependent on class and sex. Namely, both have the most impact on the estimate.

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.

```
titan_glm_inter <- glm(Survived~Age*PClass*Sex, data=data_titan)
anova(titan_glm_inter, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: gaussian, link: identity
##
## Response: Survived
##
## Terms added sequentially (first to last)
##
##
##                Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                             755        183
## Age             1      0.7       754        183    0.026 *
## PClass          2     26.0       752        157  < 2e-16 ***
## Sex             1     43.7       751        113  < 2e-16 ***
## Age:PClass      2      1.1       749        112    0.022 *
## Age:Sex         1      4.0       748        108  7.2e-08 ***
## PClass:Sex      2      4.1       746        104  4.2e-07 ***
## Age:PClass:Sex  2      0.2       744        104    0.510
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the above P-values for all possible interaction models, it seems that the Survived~Age*Sex model could give the best results. ??

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).

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).

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.

Exercise 3

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

```
data_africa <- read.table(file="data/africa.txt", header=TRUE)

# data_africa$pollib <- as.factor(data_africa$pollib)

africa_glm <- glm(miltcoup~oligarchy+pollib+parties+pctvote+popn+size+numelec+numregim, family=poisson,

summary(africa_glm)
```

```
##
## Call:
## glm(formula = miltcoup ~ oligarchy + pollib + parties + pctvote +
##     popn + size + numelec + numregim, family = poisson, data = data_africa)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.344  -0.954  -0.259   0.391   1.695
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.510269   0.905330   -0.56   0.5730
## oligarchy    0.073081   0.034596    2.11   0.0346 *
## pollib      -0.712978   0.272563   -2.62   0.0089 **
## parties      0.030774   0.011187    2.75   0.0059 **
## pctvote      0.013872   0.009753    1.42   0.1549
## popn         0.009343   0.006595    1.42   0.1566
## size        -0.000190   0.000248   -0.76   0.4445
## numelec     -0.016078   0.065484   -0.25   0.8060
## numregim     0.191735   0.229289    0.84   0.4030
## ---
## 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: 28.668  on 27  degrees of freedom
## AIC: 111.5
##
## Number of Fisher Scoring iterations: 6
```

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, oligarchy, pollib and parties show a <0.05 P-value.

? Do we need to initialize pollib as a factor variable ?

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).

```r
africa_glm <- glm(miltcoup~oligarchy+pollib+parties, family=poisson, data=data_africa)

summary(africa_glm)
```

```
##
## Call:
## glm(formula = miltcoup ~ oligarchy + pollib + parties, family = poisson,
##     data = data_africa)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.358  -1.042  -0.286   0.628   1.752
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.25138    0.37269    0.67    0.500
## oligarchy    0.09262    0.02178    4.25  2.1e-05 ***
## pollib      -0.57410    0.20438   -2.81    0.005 **
## parties      0.02206    0.00896    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: 32.856  on 32  degrees of freedom
## AIC: 105.7
##
## Number of Fisher Scoring iterations: 5
```

Delete variable with highest insignificant p-value till model only contains significant variables: Delete numelec delete numregim delete size delete popn delete pctvote Done, model now only contains oligarchy, pollib and parties as significant explanatory variables with P-values < 0.05. These are the same variables that were also already significant with subquestion 3a.

If we have to factorize then also ignore pollib1 because of insignificance.

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.

```r
Y = exp(-0.51+0.07-0.71+0.03+0.01+0.01-0.02+0.19); Y # Without factorization
```

```
## [1] 0.395
```

```r
Y_2 = exp(-0.23+0.07-1.10-1.69+0.03+0.02+0.01+0-0.3+0.21); Y_2 # with factorization?
```

```
## [1] 0.0508
```