EDDA - Assignment 3 - Group 77

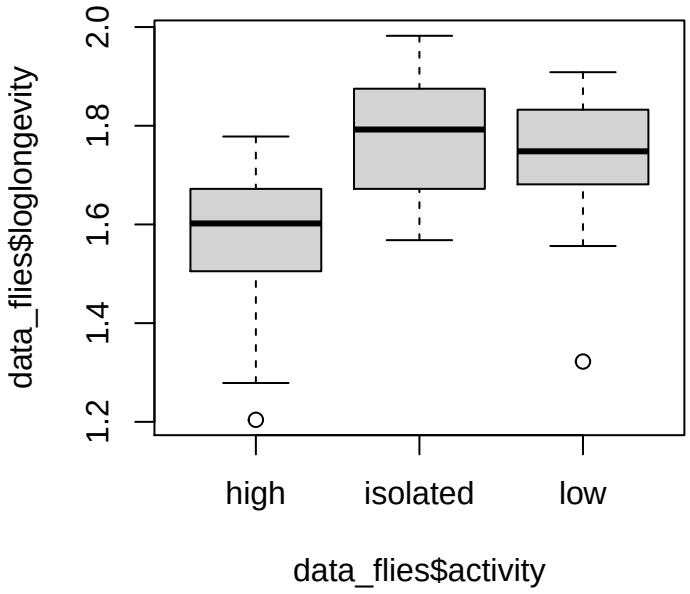
Dante de Lang, Ignas Krikštaponis and Kamiel Gülpen

Exercise 1

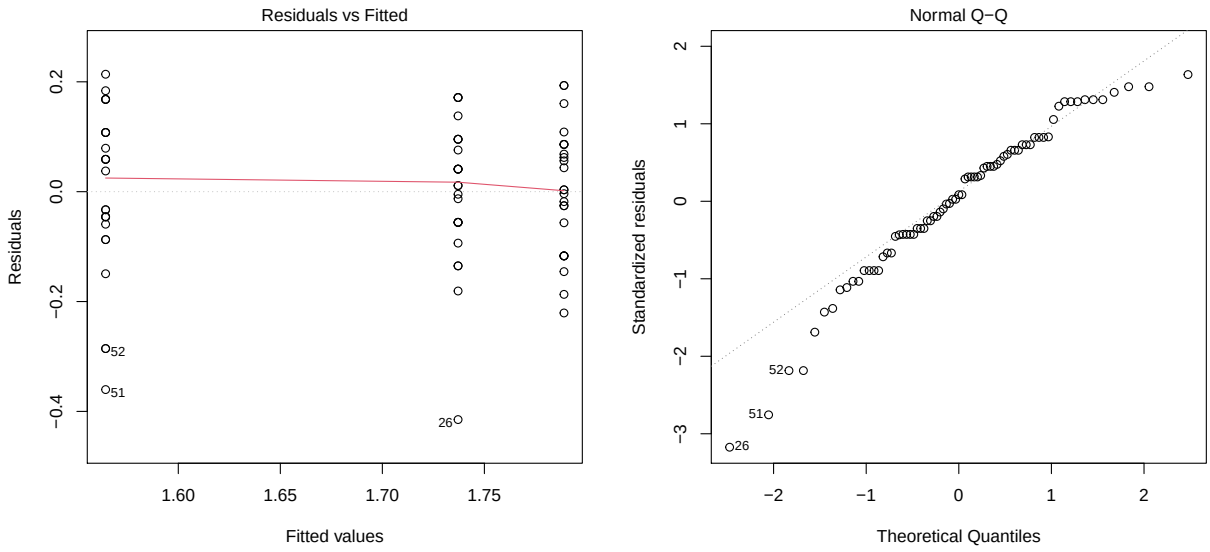
To investigate the effect of sexual activity on longevity of fruit flies, 75 male fruit flies were divided randomly in three groups of 25. The fruit flies in the first group were kept solitary, those in the second were kept together with one virgin female fruit fly per day, and those in the third group 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 = log10(longevity))
# make informative plot - boxplot
plot(data_flies$loglongevity~data_flies$activity)
```




```
# 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
```




```
# perform one-way ANOVA
anova(model); summary(model)$coefficients
```

```
## Analysis of Variance Table
```

```
##
## Response: loglongevity
##           Df Sum Sq Mean Sq F value  Pr(>F)
## activity   2  0.692   0.346    19.4 1.8e-07 ***
## Residuals 72  1.282   0.018
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

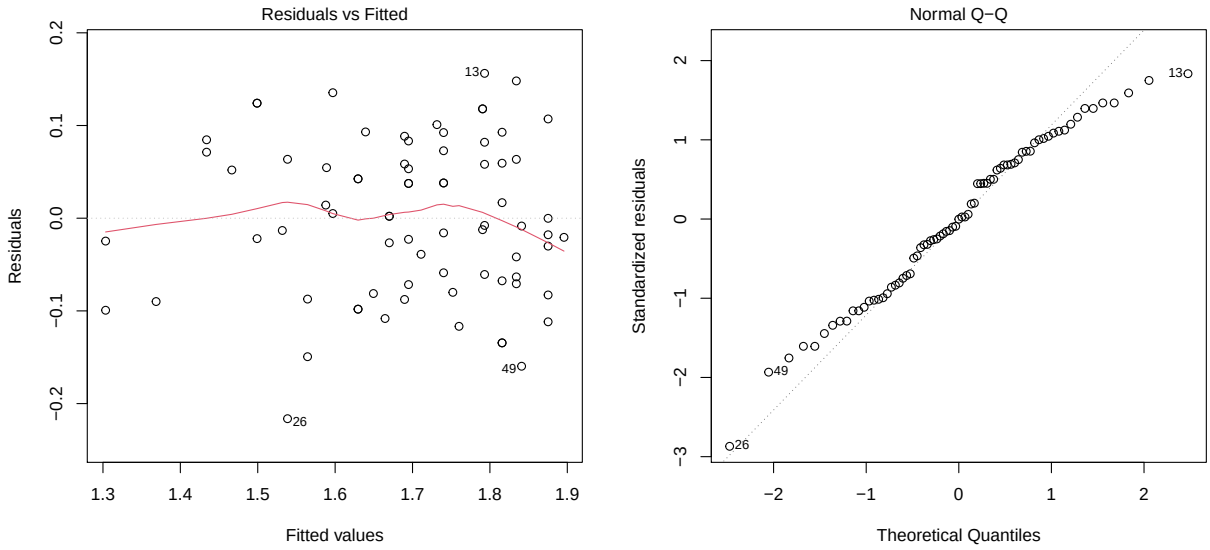
```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.564     0.0267   58.62 1.65e-62
## activityisolated     0.225     0.0377    5.95 8.82e-08
## activitylow          0.173     0.0377    4.58 1.93e-05
```

One-way ANOVA was performed to investigate whether sexual activity has an effect on loglongevity. From the results we can see that the p-values < 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 1.5644, for isolated it is $1.5644 + 0.2246 = 1.7890$ and low it is $1.5644 + 0.1727 = 1.7371$.

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?

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



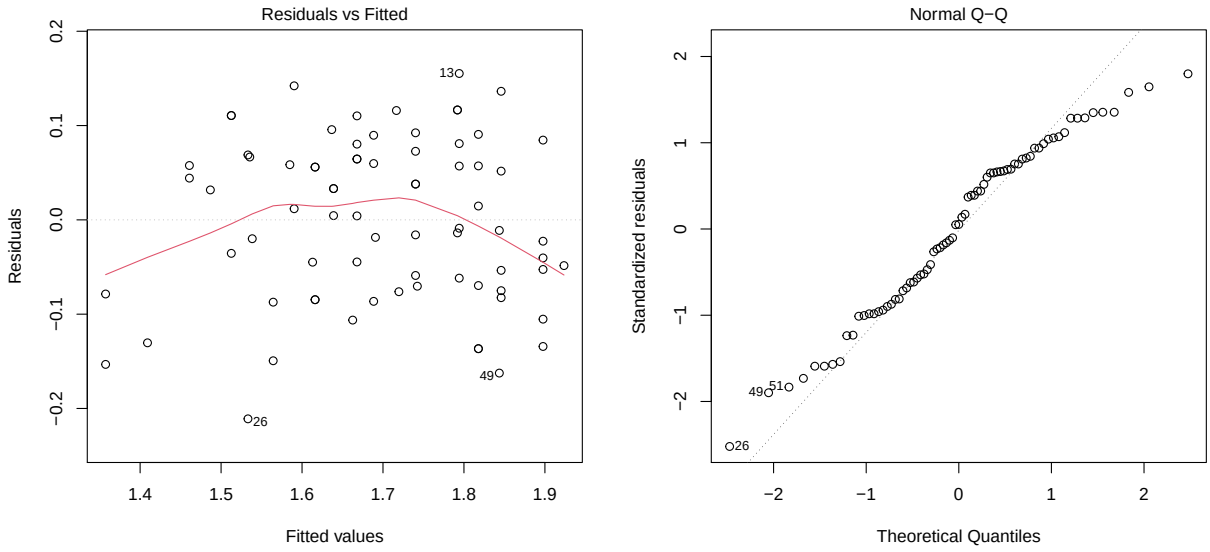

```
anova(model_interaction)
```

```
## Analysis of Variance Table
```

```
##
## Response: loglongevity
##                 Df Sum Sq Mean Sq F value  Pr(>F)
## thorax           1  1.025   1.025  135.62 < 2e-16 ***
## activity         2  0.399   0.199   26.38 3.1e-09 ***
## thorax:activity  2  0.029   0.015    1.93    0.15
## Residuals       69  0.521   0.008
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

First we investigate if there is any significant interaction between thorax and sexual activity level by performing ANCOVA with interaction (diagnostics show that data follows the required assumptions). From the results we can see that the interaction factor is insignificant and can be ignored, therefore we can now use the additive ANCOVA model.

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




```
anova(model); table <- summary(model)$coefficients; table
```

```
## Analysis of Variance Table
##
## Response: loglongevity
##           Df Sum Sq Mean Sq F value Pr(>F)
## thorax     1  1.025   1.025   132.2 <2e-16 ***
## activity   2  0.399   0.199    25.7  4e-09 ***
## Residuals 71  0.550   0.008
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##                 Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)         0.529     0.1080    4.90 5.79e-06
## thorax              1.294     0.1332    9.71 1.14e-14
## activityisolated    0.178     0.0254    7.02 1.07e-09
## activitylow         0.124     0.0254    4.88 6.18e-06
```

```
# 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$thorax)
# calculate estimates
estimate_high <- 10**(intercept + alpha_high + beta * mean_thorax)
estimate_low <- 10**(intercept + alpha_low + beta * mean_thorax)
estimate_isolated <- 10**(intercept + alpha_isolated + beta * mean_thorax)
estimates <- c(estimate_isolated, estimate_low, estimate_high)
activity_levels <- unique(as.character(data_flies$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

Model diagnostics: 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.

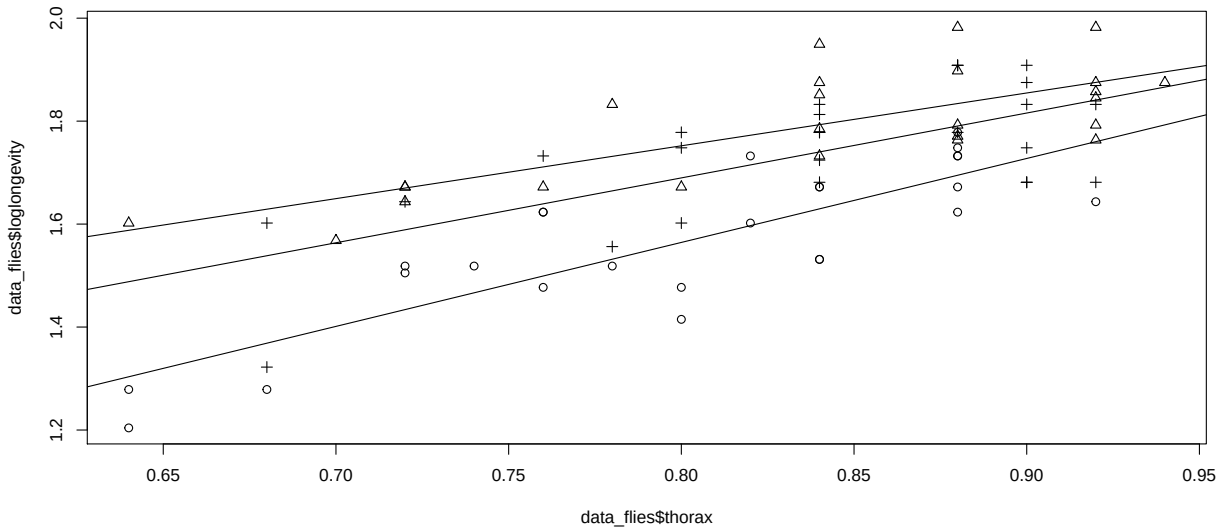
From the ANCOVA analysis results above, we can see that sexual activity has a significant effect (p-values < 0.05) on the loglongevity. From the estimates in the summary table, we can see that sexual activity decreases longevity of the fruit flies - the estimates from isolated and low sexual activity 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 there 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. 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 above it is not obvious if this is the case, the slopes look very similar. To concretely say if the slopes are the same we need to perform an ANCOVA analysis with interaction. This was already done in b) where we concluded that there is no significant interaction between thorax and sexual activity level, therefore the slope parameter can be regarded as the same between sexual activity levels.

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

All the diagnostics were performed in b)

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
```

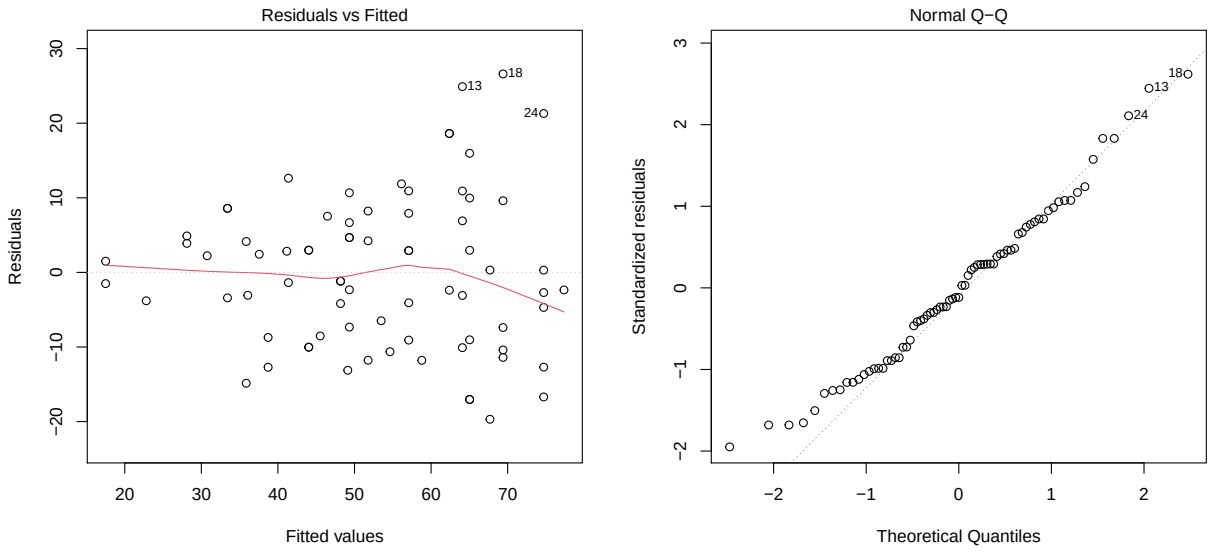
```
# extract model's parameter
intercept <- table[,1][1]; beta <- table[,1][2]
alpha_high <- 0; alpha_low <- table[,1][4]
alpha_isolated <- table[,1][3]
# calculate mean thorax
mean_thorax <- mean(data_flies$thorax)
# calculate estimates
estimate_high <- intercept + alpha_high + beta * mean_thorax
estimate_low <- intercept + alpha_low + beta * mean_thorax
estimate_isolated <- intercept + alpha_isolated + beta * mean_thorax
estimates <- c(estimate_isolated, estimate_low, estimate_high)
activity_levels <- unique(as.character(data_flies$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 a wise choice to use the logarithmic response.

Exercise 2

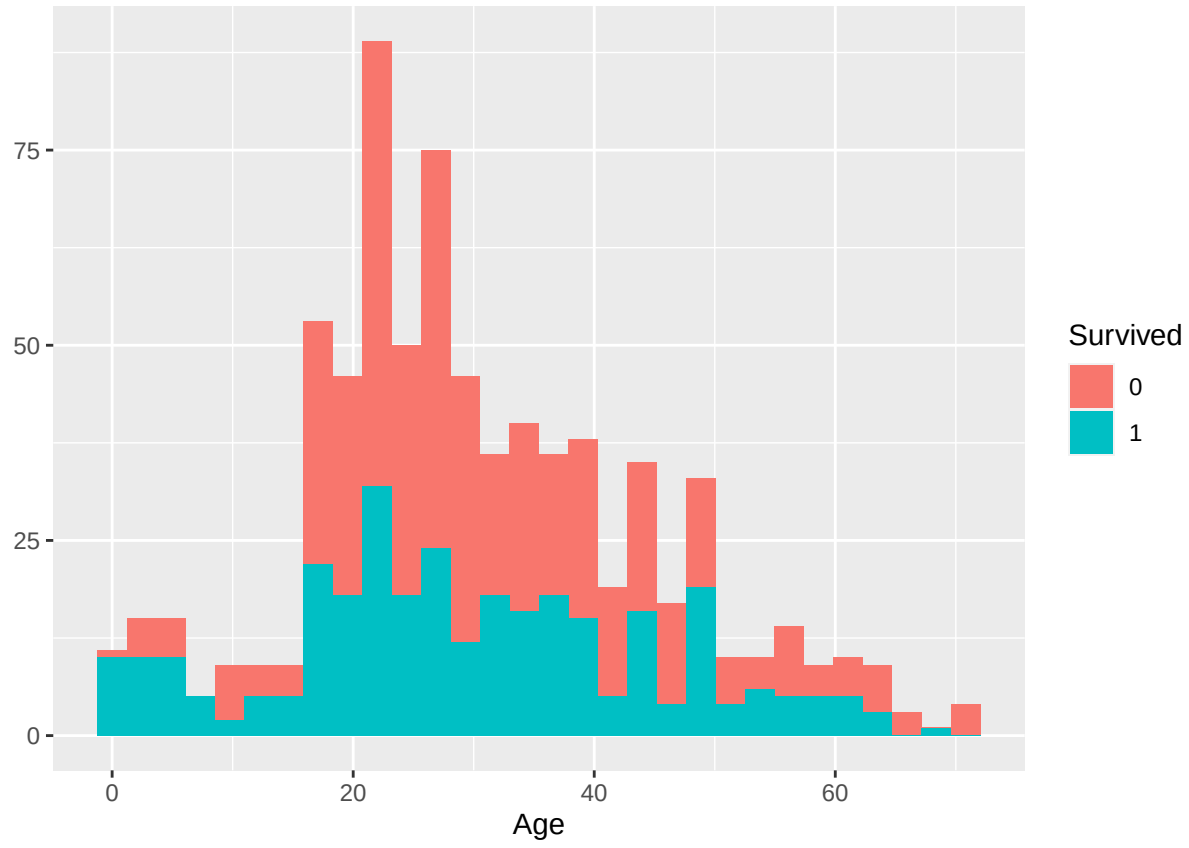
On April 15, 1912, British passenger liner Titanic sank after colliding with an iceberg. There were not enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew. The data file titanic.txt gives the survival status of passengers on the Titanic, together with 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).

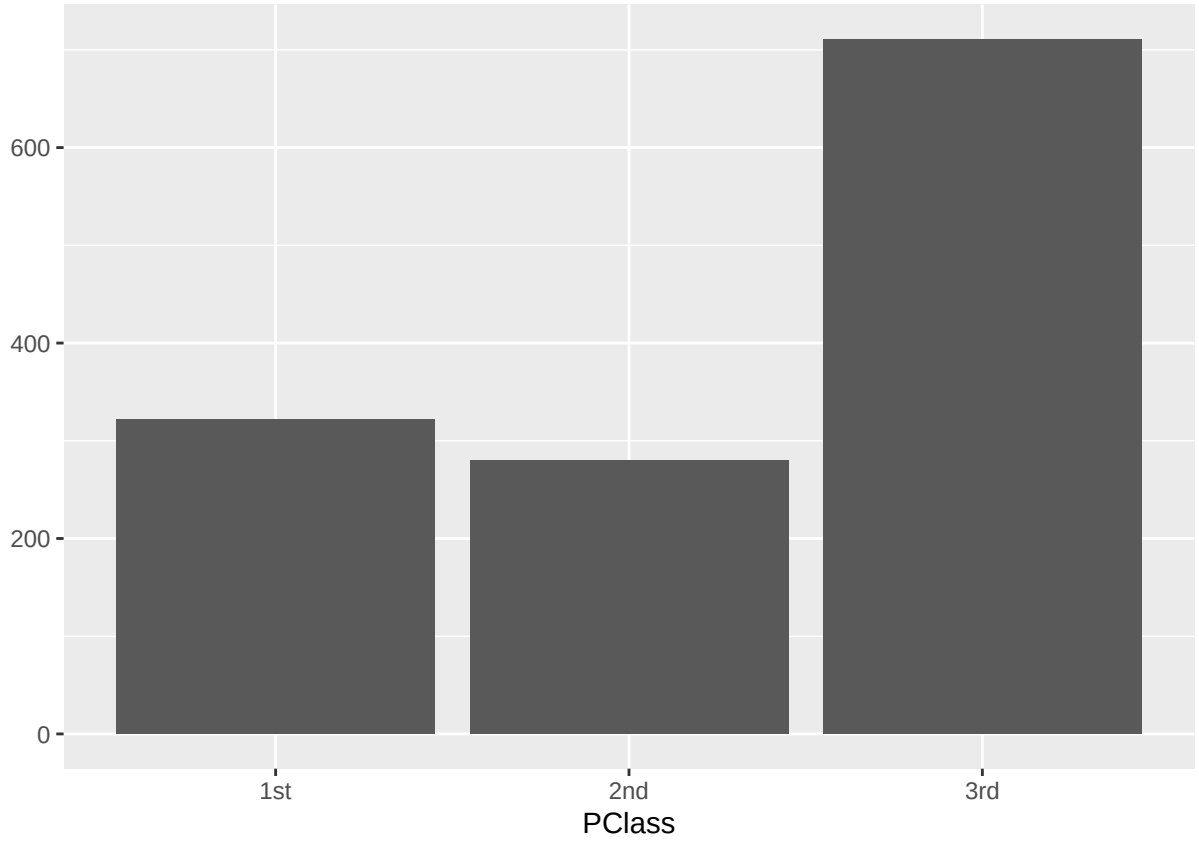
```
titanic <- read.table(file="data/titanic.txt", header=TRUE)
titanic$Survived <- as.factor(titanic$Survived)
qplot(x = Age, fill = Survived, data = titanic)
```

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

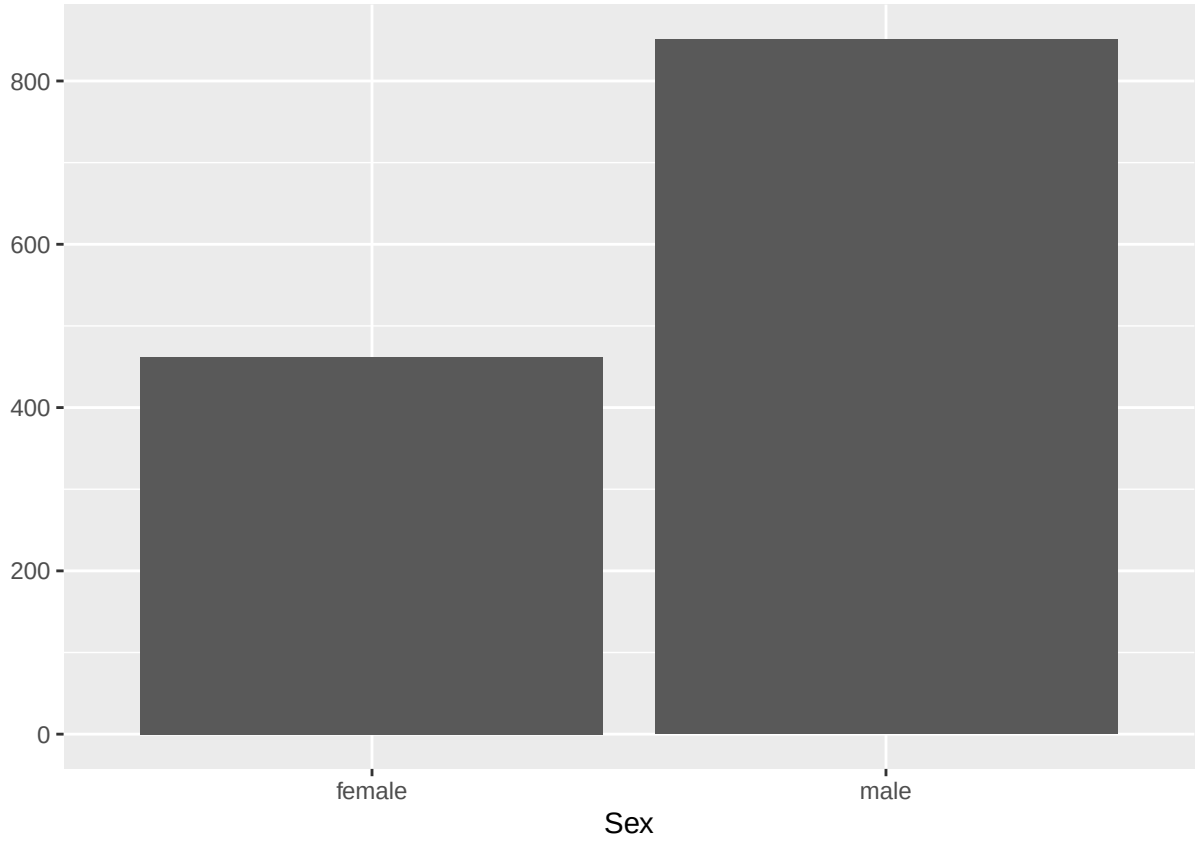
```
## Warning: Removed 557 rows containing non-finite values (stat_bin).
```



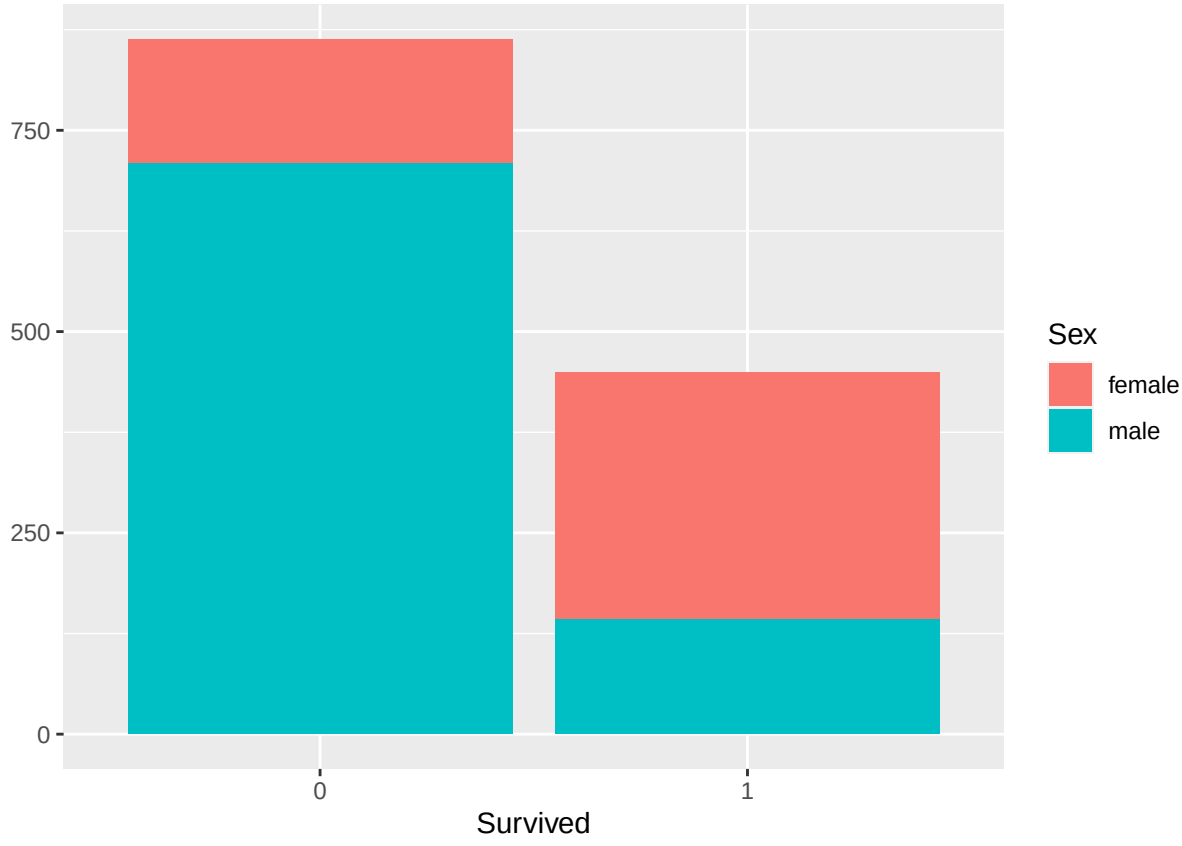
```
qplot(x = PClass, data = titanic)
```



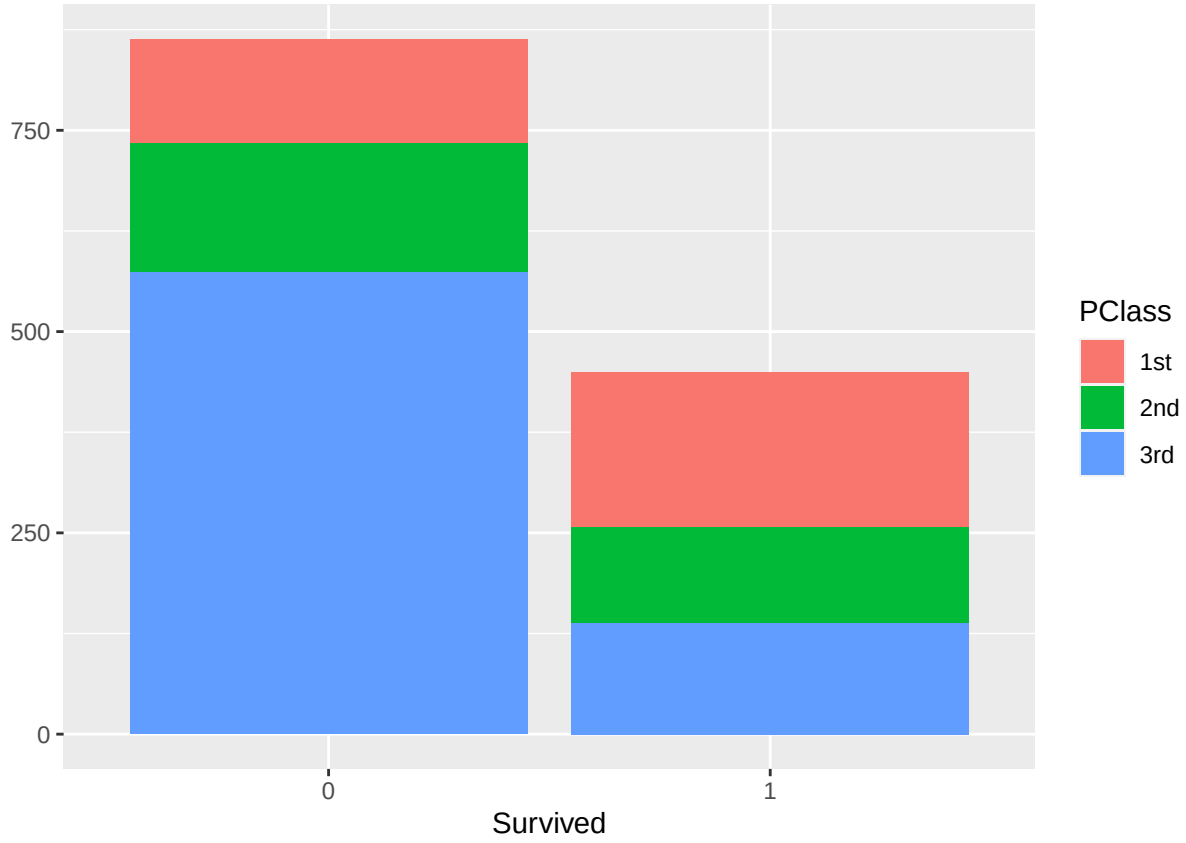

```
qplot(x = Sex, data = titanic)
```




```
qplot(x = Survived, fill = Sex, data = titanic)
```

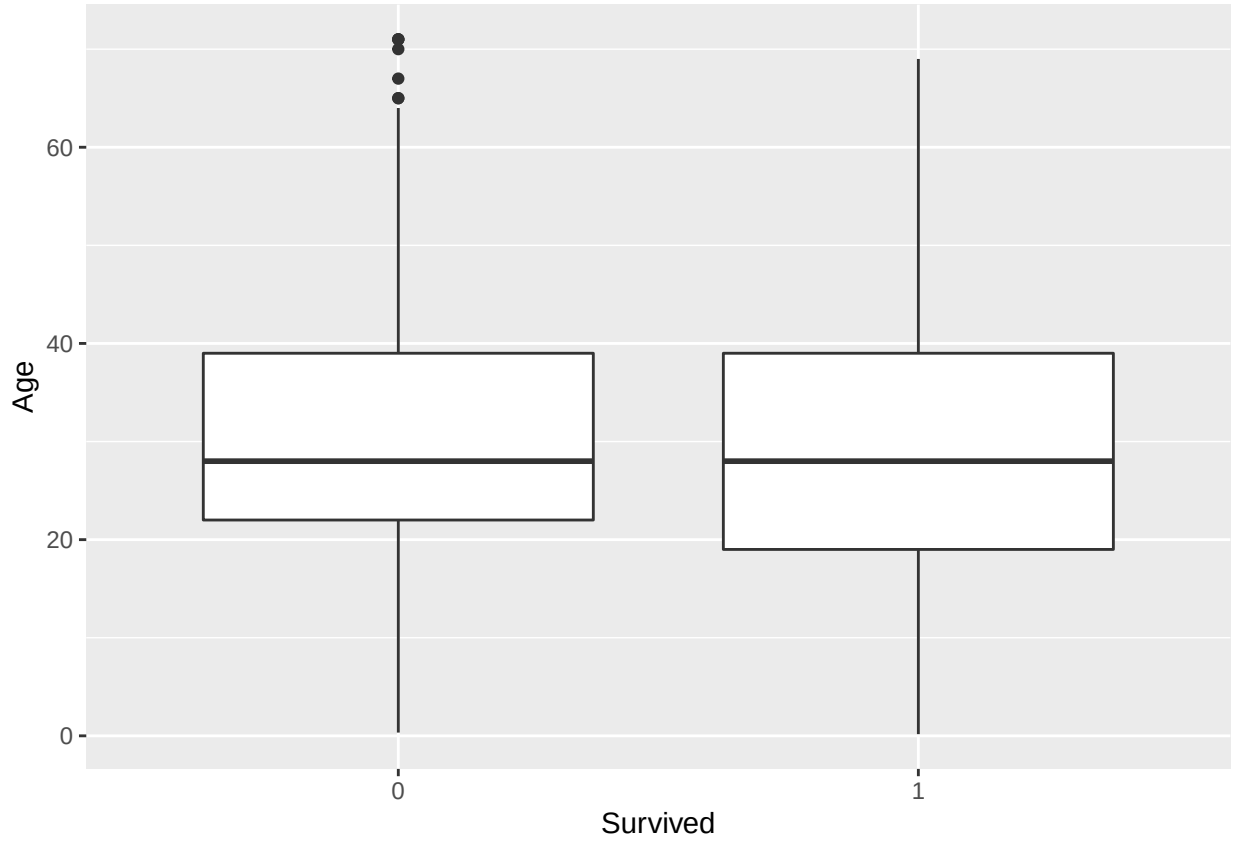



```
qplot(x = Survived, fill = PClass, data = titanic)
```




```
qplot(x = Survived, y = Age, geom = "boxplot", data = titanic)
```

```
## Warning: Removed 557 rows containing non-finite values (stat_boxplot).
```



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.

```
# what to do with NAs?
titanic$PClass <- as.factor(titanic$PClass)
titanic$Sex <- as.factor(titanic$Sex)
logistic <- glm(Survived~PClass+Age+Sex, data = titanic, family = binomial)
summary(logistic)
```

```
##
## Call:
## glm(formula = Survived ~ PClass + Age + Sex, family = binomial,
##     data = titanic)
##
## 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
```

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. d) Propose a method to predict the survival status and a quality measure for your prediction

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

```
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)               2.5105     1.1856  2.1174   0.0342
## PClass2nd                 0.3693     1.4985  0.2465   0.8053
## PClass3rd                -2.6915     1.2595 -2.1370   0.0326
## Age                       0.0122     0.0309  0.3930   0.6943
## Sexmale                  -1.0055     1.3338 -0.7538   0.4510
## PClass2nd:Age            -0.0419     0.0415 -1.0091   0.3129
## PClass3rd:Age            -0.0128     0.0351 -0.3665   0.7140
## PClass2nd:Sexmale        -0.1458     1.7713 -0.0823   0.9344
## PClass3rd:Sexmale         0.6815     1.4873  0.4582   0.6468
## Age:Sexmale              -0.0664     0.0344 -1.9306   0.0535
## PClass2nd:Age:Sexmale    -0.0478     0.0543 -0.8791   0.3793
## PClass3rd:Age:Sexmale     0.0164     0.0432  0.3810   0.7032
```

```
# predict for 53 years and all PClass
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, new_data, type="response")
final <- new_data %>% bind_cols(Survival = results)
knitr::kable(final)
```

PClass	Sex	Age	Survival
1st	female	53	0.843
2nd	female	53	0.597
3rd	female	53	0.302
1st	male	53	0.279
2nd	male	53	0.096
3rd	male	53	0.030

From the model above we do not see any significant interaction between predictors, therefore we choose the model from *b*).

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. 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).

```
new_model <- glm(Survived~PClass+Sex, data = titanic, family = binomial)
drop1(new_model,test="Chisq")
```

```
## Single term deletions
##
## Model:
## Survived ~ PClass + Sex
##        Df Deviance  AIC LRT Pr(>Chi)
## <none>        1199 1207
## PClass  2     1356 1360 156   <2e-16 ***
## Sex     1     1515 1521 316   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the test above we can conclude 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.

#Exercise 3

To study the in uence 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=poisso
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
```

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).

```
summary(glm(miltcoup~oligarchy+pollib+parties+pctvote+popn+size+numelec+numregim,family=poisson,data=af
```

```
##
## 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
```

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

```
##
## Call:
## glm(formula = miltcoup ~ oligarchy + pollib + parties + pctvote +
##     popn + size + numregim, family = poisson, data = africa)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.535  -0.941  -0.313   0.424   1.664
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.457746   0.860234   -0.53   0.5946
## oligarchy    0.081202   0.028815    2.82   0.0048 **
## pollib1     -0.964298   0.562094   -1.72   0.0862 .
## pollib2     -1.514951   0.526944   -2.87   0.0040 **
## parties      0.029341   0.010310    2.85   0.0044 **
## pctvote      0.013912   0.009465    1.47   0.1416
## popn         0.009959   0.006725    1.48   0.1386
## size        -0.000269   0.000269   -1.00   0.3171
## numregim     0.180442   0.224117    0.81   0.4207
## ---
## 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: 28.430  on 27  degrees of freedom
## AIC: 111.2
##
## Number of Fisher Scoring iterations: 5
```

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

```
##
## Call:
## glm(formula = miltcoup ~ oligarchy + pollib + parties + pctvote +
##     popn + size, family = poisson, data = africa)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.551  -0.896  -0.223   0.526   1.606
```

```
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.041976   0.577410    0.07  0.94205
## oligarchy    0.089495   0.027044    3.31  0.00094 ***
## pollib1     -0.967325   0.560560   -1.73  0.08441 .
## pollib2     -1.532113   0.523278   -2.93  0.00341 **
## parties      0.028817   0.010217    2.82  0.00480 **
## pctvote      0.014922   0.009376    1.59  0.11151
## popn         0.007165   0.005684    1.26  0.20751
## size        -0.000258   0.000266   -0.97  0.33262
## ---
## 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: 29.081  on 28  degrees of freedom
## AIC: 109.9
##
## Number of Fisher Scoring iterations: 5
```

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

```
##
## Call:
## glm(formula = miltcoup ~ oligarchy + pollib + parties + pctvote +
##     popn, family = poisson, data = africa)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.420  -0.995  -0.144   0.570   1.611
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.23143    0.52889   -0.44   0.6617
## oligarchy    0.08347    0.02583    3.23   0.0012 **
## pollib1     -0.68359    0.49582   -1.38   0.1680
## pollib2     -1.32057    0.49027   -2.69   0.0071 **
## parties      0.02977    0.01031    2.89   0.0039 **
## pctvote      0.01392    0.00937    1.49   0.1373
## popn         0.00566    0.00548    1.03   0.3020
## ---
## 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: 30.040  on 29  degrees of freedom
## AIC: 108.8
##
## Number of Fisher Scoring iterations: 5
```

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

```
##
## Call:
## glm(formula = miltcoup ~ oligarchy + pollib + parties + pctvote,
##     family = poisson, data = africa)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.530  -0.979  -0.183   0.566   1.672
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.11650    0.51375   -0.23   0.8206
## oligarchy    0.09471    0.02318    4.09  4.4e-05 ***
## pollib1     -0.62076    0.48753   -1.27   0.2029
## pollib2     -1.31037    0.48902   -2.68   0.0074 **
## parties      0.02574    0.00955    2.70   0.0070 **
## pctvote      0.01206    0.00907    1.33   0.1838
## ---
## 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: 31.069  on 30  degrees of freedom
## AIC: 107.9
##
## Number of Fisher Scoring iterations: 5
```

```r
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
```

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;
avg_pctvote <- new_data$pctvote; avg_popn <- new_data$popn;
avg_size <- new_data$size; avg_numelec <- new_data$numelec;
avg_numregim <- new_data$numregim
# get a list of polical liberalization levels
pollib <- as.character(unique(africa$pollib))
new_data <- expand.grid(oligarchy = avg_oligarchy, pollib = pollib,
                        parties = avg_parties, pctvote = avg_pctvote,
                        popn = avg_popn, size = avg_size,
                        numelec = avg_numelec, numregim = avg_numregim)
# full model
results <- predict(poisson_model, new_data, type="response")
final <- new_data %>% bind_cols(Prediction = results)
knitr::kable(final, caption = "Predictions with the full model")
```

Table 4: Predictions with the full model

oligarchy	pollib	parties	pctvote	popn	size	numelec	numregim	Prediction
5.22	1	17.1	32.1	11.6	485	6.72	2.75	1.570
5.22	2	17.1	32.1	11.6	485	6.72	2.75	0.873
5.22	0	17.1	32.1	11.6	485	6.72	2.75	4.731

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

Table 5: Predictions with the reduced model

oligarchy	pollib	parties	pctvote	popn	size	numelec	numregim	Prediction
5.22	1	17.1	32.1	11.6	485	6.72	2.75	1.772
5.22	2	17.1	32.1	11.6	485	6.72	2.75	0.956
5.22	0	17.1	32.1	11.6	485	6.72	2.75	2.908