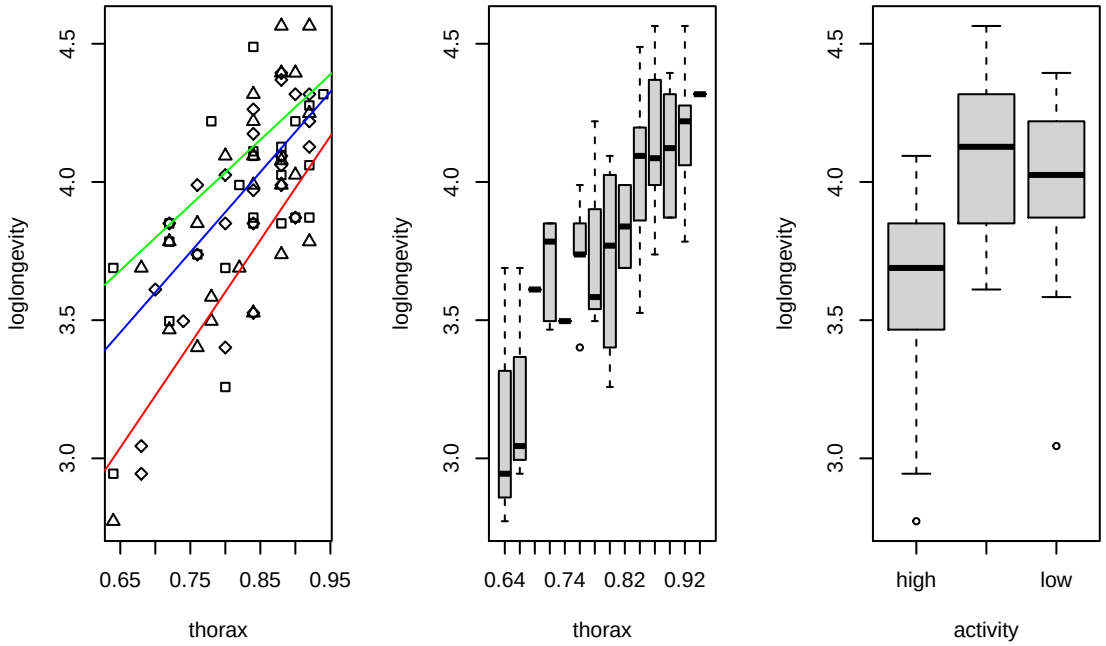
Assignment 3

Andrei Udriste, Xinyu Hu, Maria Gherghina-Tudor - Group 43

2021/3/21

Exercise 1.

a)



It is quite clear there exists a positive linear relationship between loglongevity and thorax. And the boxplots shows that loglongevity has influenced by thorax and activity.

```
ffaov = lm(loglongevity~activity, data=ffdata);anova(ffaov)
```

```
## Analysis of Variance Table
##
## Response: loglongevity
##           Df Sum Sq Mean Sq F value    Pr(>F)
## activity   2 3.6665  1.8333  19.421 1.798e-07 ***
## Residuals 72 6.7966  0.0944
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The anova test for testing H_0 shows the p-value = 1.798e-07 < 0.05, which concludes longevity is effected by activity.

```
summary(ffaov)
```

```
##
## Call:
## lm(formula = loglongevity ~ activity, data = ffdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95531 -0.13338  0.02552  0.20891  0.49222
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.60212    0.06145  58.621  < 2e-16 ***
## activityisolated  0.51722    0.08690   5.952 8.82e-08 ***
## activitylow       0.39771    0.08690   4.577 1.93e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3072 on 72 degrees of freedom
## Multiple R-squared:  0.3504, Adjusted R-squared:  0.3324
## F-statistic: 19.42 on 2 and 72 DF,  p-value: 1.798e-07
```

From above summary, $\hat{\mu} = 3.60212$, $\hat{\alpha}_{isolated}$=0.51722 and $\hat{\alpha}_{low}$=0.39771. Thus, estimated longevity of high $e^{3.60212}$=36.67, of isolated $e^{3.60212+0.51722}$=61.51 and of low $e^{3.60212+0.39771}$=54.58.

b)

```
fflm = lm(loglongevity~thorax+activity,data=ffdata);drop1(fflm,test="F")
```

```
## Single term deletions
##
## Model:
## loglongevity ~ thorax + activity
##          Df Sum of Sq    RSS     AIC F value    Pr(>F)
## <none>                2.9180 -235.50
## thorax    1    3.8786 6.7966 -174.08  94.374 1.139e-14 ***
## activity  2    2.1129 5.0309 -198.64  25.705 4.000e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With Single term deletions, H_0 is rejected because of p-value=4.000e-09 < 0.05. It shows that activity still influences longevity.

```
summary(fflm)
```

```
##
## Call:
## lm(formula = loglongevity ~ thorax + activity, data = ffdata)
##
## 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)        1.21893    0.24865   4.902 5.79e-06 ***
## thorax             2.97899    0.30665   9.715 1.14e-14 ***
## activityisolated   0.40998    0.05839   7.021 1.07e-09 ***
## activitylow        0.28570    0.05849   4.885 6.18e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2027 on 71 degrees of freedom
## Multiple R-squared:  0.7211, Adjusted R-squared:  0.7093
## F-statistic:  61.2 on 3 and 71 DF,  p-value: < 2.2e-16
```

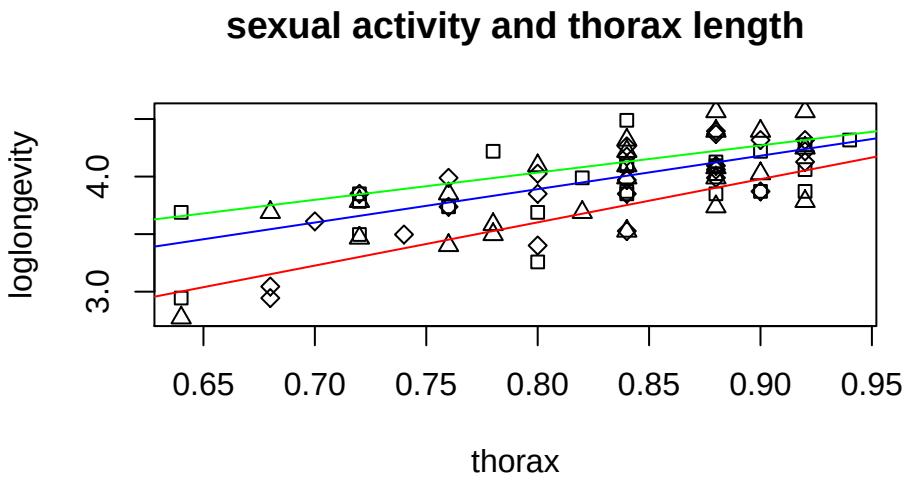
The summary shows the $\hat{\mu}$=1.21893, $\hat{\alpha}_{isolated}$=0.40998 and $\hat{\alpha}_{low}$=0.28570. In this case, activity decreases longevity.

```
mean(ffdata$thorax)
```

```
## [1] 0.8245333
```

The average thorax length = 0.8245333. Since the model is $\hat{Y}_i = \hat{\mu} + \hat{\alpha}_i + \hat{\beta} X_i$, $\hat{Y}_{isolated} = 1.21893 + 0.40998 + 2.97899 * 0.8245333 = 4.085186455367$, $\hat{Y}_{low} = 1.21893 + 0.28570 + 2.97899 * 0.8245333 = 3.960906455367$ and $\hat{Y}_{high} = 1.21893 + 0 + 2.97899 * 0.8245333 = 3.675206455367$. Additionally, the longevity of isolated $e^{Y_{isolated}} = e^{4.085186455367}$=59.45, of low $e^{Y_{low}} = e^{3.960906455367}$=52.50 and of high $e^{Y_{high}} = e^{3.675206455367}$=39.45.

c)




```
fflmm=lm(loglongevity~activity*thorax,data=ffdata);summary(fflmm)
```

```
##
## Call:
## lm(formula = loglongevity ~ activity * thorax, data = ffdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49803 -0.15920 -0.00031  0.14624  0.35984
##
```

```
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.5978     0.4192   1.426   0.1584
## activityisolated           1.5465     0.5845   2.646   0.0101 *
## activitylow                0.9717     0.6423   1.513   0.1349
## thorax                     3.7554     0.5216   7.199 5.78e-10 ***
## activityisolated:thorax   -1.3929     0.7122  -1.956   0.0545 .
## activitylow:thorax        -0.8539     0.7794  -1.096   0.2771
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2001 on 69 degrees of freedom
## Multiple R-squared:  0.7359, Adjusted R-squared:  0.7167
## F-statistic: 38.44 on 5 and 69 DF,  p-value: < 2.2e-16
```

The graph shows that each estimate for each group depending on the thorax length, and three lines are parallel. Additionally, from the summary, it shows that p-value for $H_0 : \mu_{low} = \mu_{high}$ is $0.2771 > 0.05$ and p-value for $H_0 : \mu_{isolated} = \mu_{high}$ is $0.0545 > 0.05$. Therefore, we reject both of them, and this dependence is similar under all three conditions of sexual activity.

d)

```
summary(ffaov)$r.squared
```

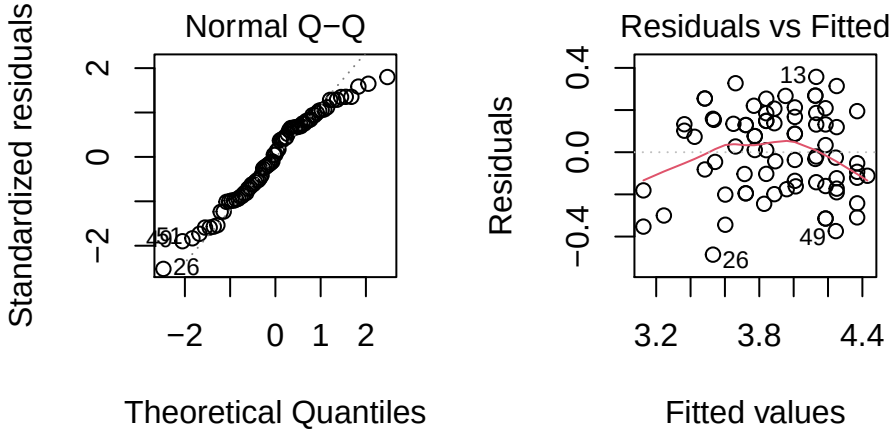
```
## [1] 0.3504222
```

```
summary(fflm)$r.squared
```

```
## [1] 0.721116
```

The analyses with thorax length is preferred, because the one with thorax length has a explained variance of 70.9%, but the one without thorax length has only 35%. None of them is wrong.

e)




```
mean(ffdata$thorax)
```

```
## [1] 0.8245333
```

```
shapiro.test(residuals(fflm))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fflm)
## W = 0.96838, p-value = 0.05748
```

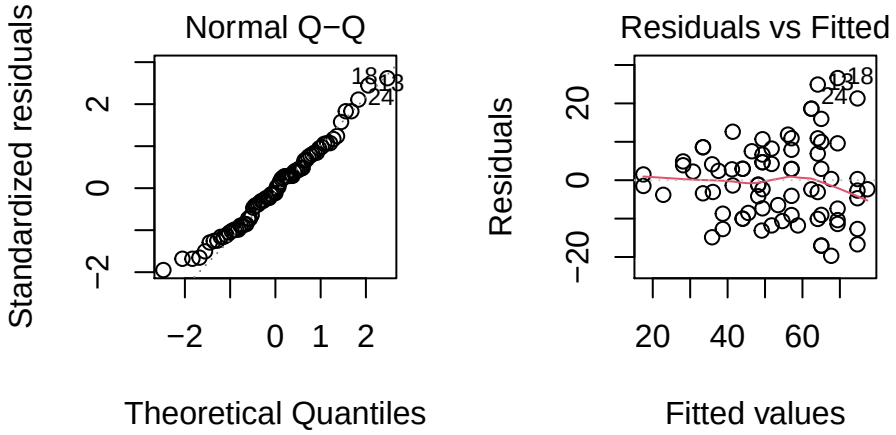
By checking QQ-plot and residuals vs. fitted plot, ANCOVA with thorax length and activity are independent to each other. In addition, we did not find certain patterns exist in the graph, even though they look normally distributed which is also checked by Shapiro-Wilk normality test.

f)

```
fflaov=lm(longevity~thorax+activity, data=ffdata)
anova(fflaov)
```

```
## Analysis of Variance Table
##
## Response: longevity
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## thorax     1 10959.3 10959.3 101.409 2.557e-15 ***
## activity   2  4966.7  2483.4  22.979 2.016e-08 ***
## Residuals 71  7673.0   108.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Even without the logarithm, the ANCOVA shows very similar with the one with logarithm.



The graphs show that the residuals are normally distributed. And we can see a certain pattern existing in residuals fitted plot. This is also supported by Shapiro-Wilk normality test below.

```
shapiro.test(residuals(fflaov))
```

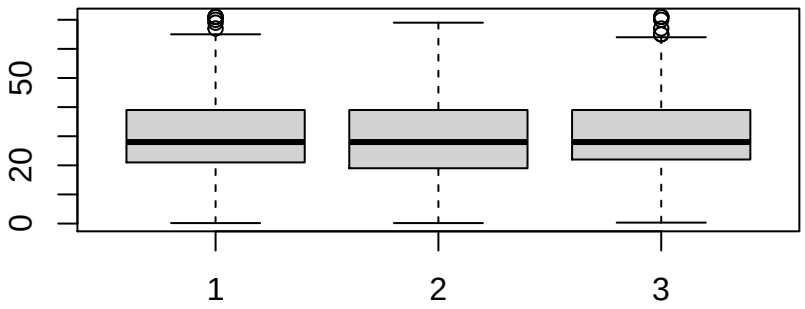
```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  residuals(fflaov)
## W = 0.98091, p-value = 0.3176
```

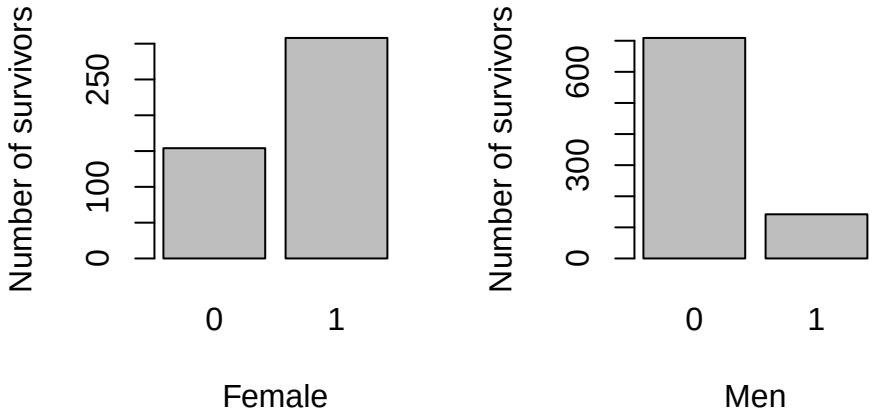
However, the variance of residuals is getting larger as the the estimates gets larger. It does not reliable. Moreover, it is wise to use the model with logarithm.

Exercise 2.

a)

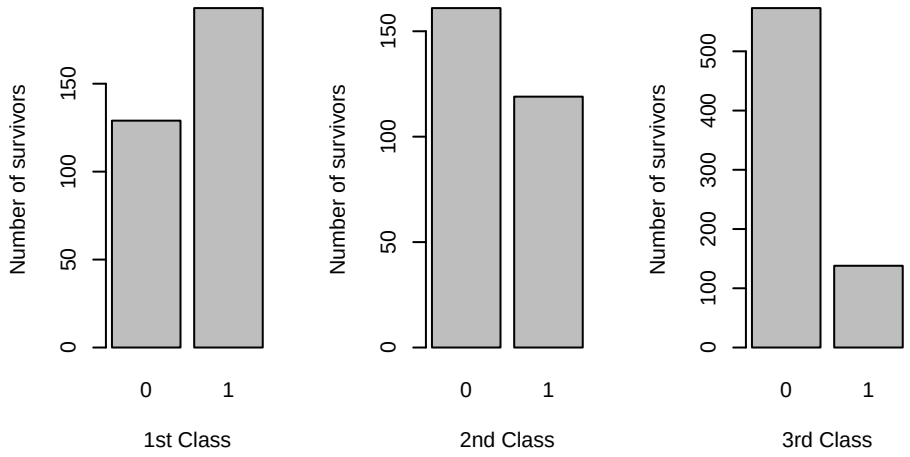


In boxplot (1) we have the age of all the passengers, (2) represent the ages of all surviving passengers and in (3) we have the ages of the passengers that didn't survive. It can be observed from the boxplot above tat the age didn't have any major impact in someone surviving or not.



We can observe a clear difference in between the two barplots, to be more precise the number of female survivors is much more higher than the one of male survivors. Contrary in the case of non-surviving passengers

the number of females is smaller than the number of males.



More differences can be observed in the survival rates of the passengers from the 3 classes. In the 1 class we can see that the number of survivors is higher than the number of non-survivors, unfortunately we can not say the same for the other 2 classes where the number non-survivors is higher than the number of survivors. This can be observed mostly in the 3rd class where the number of non-survivors is 4 times bigger than the number of survivors. One common thing that can be observed for all classes is that the number of survivors is between 100 and 200 for all 3 of them.

b)

```
fact_PClass = factor(PClass)
fact_Sex = factor(Sex)
fact_Age = factor(Age)
titglm = glm(Survived ~ fact_PClass + fact_Sex + fact_Age, data = data, family=binomial)
#summary(titglm, maxsum = 10)
```

Because of the big number of variable that appear in the summary is recommended to use Age as a numerical variable rather than a factorial one, but this change will affect the model, some of the rows will not appear, because the Age is not present (NA), and the glm model will just ignore those rows.

```
titglm_f = glm(Survived ~ fact_PClass + fact_Sex + Age, data = data, family=binomial)
summary(titglm_f)
```

```
##
## Call:
## glm(formula = Survived ~ fact_PClass + fact_Sex + Age, family = binomial,
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7226  -0.7065  -0.3917   0.6495   2.5289
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     3.759662   0.397567   9.457  < 2e-16 ***
```

```
## fact_PClass2nd -1.291962   0.260076  -4.968 6.78e-07 ***
## fact_PClass3rd -2.521419   0.276657  -9.114  < 2e-16 ***
## fact_Sexmale   -2.631357   0.201505 -13.058  < 2e-16 ***
## Age            -0.039177   0.007616  -5.144 2.69e-07 ***
## ---
## 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: 705.14
##
## Number of Fisher Scoring iterations: 5
```

After computing the smaller model we can compute all the odds by summing all the Estimates of the factors that are of interest and applying exp(sum), which translates to e to the power of the sum.

```
exp(coef(titglm_f))
```

```
##    (Intercept) fact_PClass2nd fact_PClass3rd   fact_Sexmale            Age
##    42.93391621     0.27473112     0.08034550     0.07198073     0.96158067
```

```
## odds of 1st class + female + age = 41.28441
```

```
## odds of 1st class + male + age = 1.089128
```

```
## odds of 2st class + female + age = 11.34212
```

```
## odds of 2st class + male + age = 0.8164138
```

```
## odds of 3st class + female + age = 3.317017
```

```
## odds of 3st class + male + age = 0.2387613
```

After computing the we can observe that the same trends that where present in the graphs from point **a)** are present also here. More exactly the female have a higher odds than male and as the class is lower the odds of survival also decrease. Based on this we can see that the biggest odds of survival are for the females from the 1st class, while the worst ones are for the male in the 3rd class.

c)

```
glm1 = glm(Survived ~ Age * fact_Sex, data = data, family=binomial)
anova(glm1, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Survived
##
## Terms added sequentially (first to last)
##
##
##              Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                           755    1025.57
## Age           1    2.849       754    1022.72   0.09141 .
## fact_Sex      1  227.138       753     795.59 < 2.2e-16 ***
```

```
## Age:fact_Sex  1   25.030       752     770.56 5.645e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After verifying the interaction between the factors (Sex) and the numerical variable (Age) we obtain a p-value of 5.64e-07 which is smaller than 0.05. Because of this we can conclude that we will reject $H0$ so the numerical variable Age does have a big influence over the outcome so it is not recommended to eliminate it from the model.

```
glm2 = glm(Survived ~ Age * fact_PClass, data = data, family=binomial)
anova(glm2, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Survived
##
## Terms added sequentially (first to last)
##
##
##                 Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                             755    1025.57
## Age              1    2.849       754    1022.72  0.09141 .
## fact_PClass      2  112.807       752     909.92  < 2e-16 ***
## Age:fact_PClass  2    1.166       750     908.75  0.55816
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After verifying the interaction between the two factors (Class) and the numerical variable (Age) we obtain a p-value of 0.558 which is bigger than 0.05. Because of this we can conclude that we will not reject $H0$ so the numerical variable Age does not have a big influence over the outcome so we can eliminate it from the model.

Because of the interaction between Age and PClass we will create a model that is not dependent on Age.

Because the variable Age doesn't have major influence in the outcome we can create a model that does not include it.

```
p1m = predict(titglm2, data.frame(fact_PClass = "1st", fact_Sex = "male", Age = "53"), type="response")
p2m = predict(titglm2, data.frame(fact_PClass = "2nd", fact_Sex = "male", Age = "53"), type="response")
p3m = predict(titglm2, data.frame(fact_PClass = "3rd", fact_Sex = "male", Age = "53"), type="response")
p1f = predict(titglm2, data.frame(fact_PClass = "1st", fact_Sex = "female", Age = "53"), type="response
p2f = predict(titglm2, data.frame(fact_PClass = "2nd", fact_Sex = "female", Age = "53"), type="response
p3f = predict(titglm2, data.frame(fact_PClass = "3rd", fact_Sex = "female", Age = "53"), type="response
```

We can use the **predict()** function to predict information about a specific group of passengers. But the problem is that the predicted value is given in probability instead of odds, so we have to transform it in odds.

```
## odds of 1st class + male + age(53) = 0.6097259
```

```
## odds of 2nd class + male + age(53) = 0.2768802
```

```
## odds of 3rd class + male + age(53) = 0.0793519
```

```
## odds of 1st class + female + age(53) = 7.033281
```

```
## odds of 2nd class + female + age(53) = 3.193855
```

```
## odds of 3rd class + female + age(53) = 0.9153361
```

After transforming the probabilities in odds it can be observed that the same pattern is maintained, mainly that the odds for survival for females are higher than the ones for males and as the class increase the odds decrease.

d)

One method to predict the survival status would be to split the data in 3 categories:

- Training data
- Validation data
- Test data

THe training data can be used to create a model and to optimize it (obtain the maximum likelihood of teta_hat). After the model has been train we can try and use the validation data to predict how will the model predict the outcome. If the accuracy of the model on the validation data is as good or better than the imposed threshold we can move to the next step, test the model on the test data. This is one of the most important steps, because it can offer us a pretty good measurement of the model accuracy. It is recommended to not use the test data more than a few times (is ideal to use it only once) to test the model.

If the model offers a good accuracy on the test data then we can say that our model will perform pretty well in a real scenario.

e)

```
##          fact_PClass
## Survived 1st 2nd 3rd
##        0 129 161 573
##        1 193 119 138
```

Using the contingency table for the two factors PClass and Survived it can be observed that the trends present in the graphs presented above is maintained.

```
tot_ch = chisq.test(tot_class)
tot_ch
```

```
##
##  Pearson's Chi-squared test
##
## data:  tot_class
## X-squared = 172.3, df = 2, p-value < 2.2e-16
```

Because the p-value is equal with 2.2e-16 which is smaller then 0.05, it can be concluded that we reject the null hypotheses H0. This means that there is a dependence between the two factors PClass and Survived.

```
tot_sex = xtabs(~ Survived + fact_Sex)
tot_sex
```

```
##          fact_Sex
## Survived female male
##        0    154  709
##        1    308  142
```

Using the contingency table for the two factors Sex and Survived it can be observed that the trends present in the graphs presented above is maintained.

```
fisher.test(tot_sex)
```

```
##
##  Fisher's Exact Test for Count Data
```

```
##
## data:  tot_sex
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.07620521 0.13155709
## sample estimates:
## odds ratio
##  0.1003494
```

After applying the Fisher test we obtain a p-value equal with 2.2e-16 which is smaller then 0.05, so it can be concluded that we reject the null hypotheses H0. This means that there is a dependence between the two factors Sex and Survived.

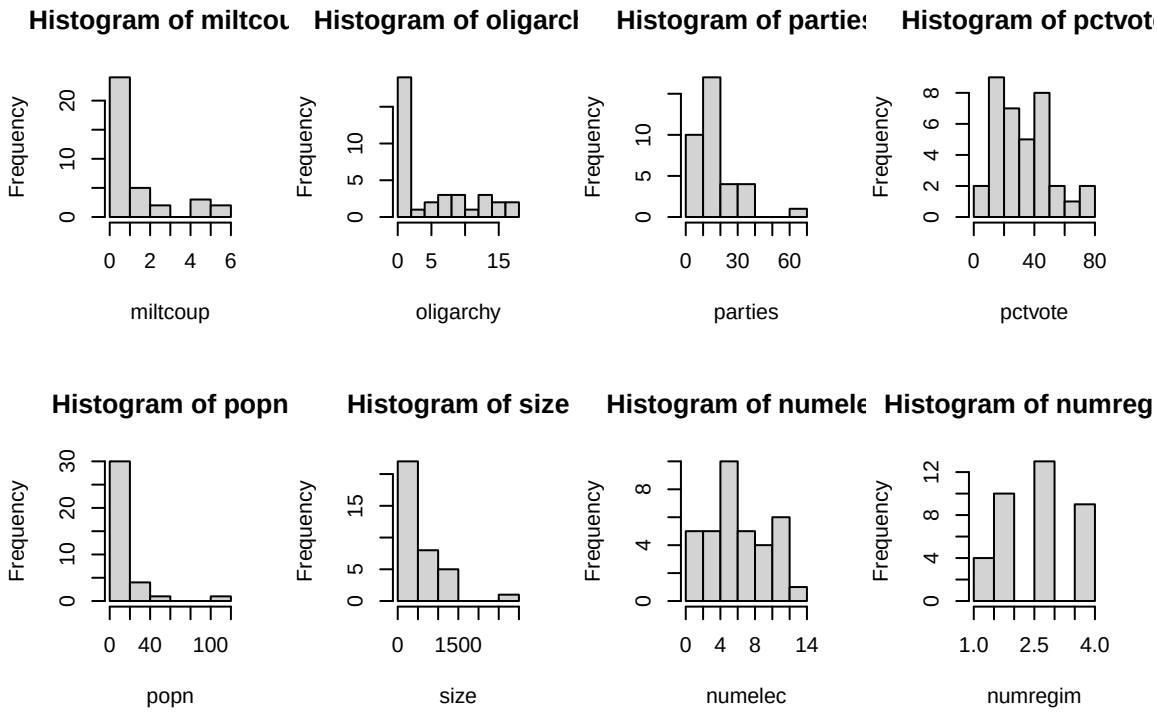
f)

Yes, the second approach is wring because it only test one factor at a time instead of both at the same time. On advantage of this approach is that we can test and see the influence of each factor separately on the Survived status.Based on this approach we can discard the factors that do not influence the outcome. One disadvantage is that we can only test one variable at at time and we can not see if there is any influence between two factors (ex: between PClass and Sex).

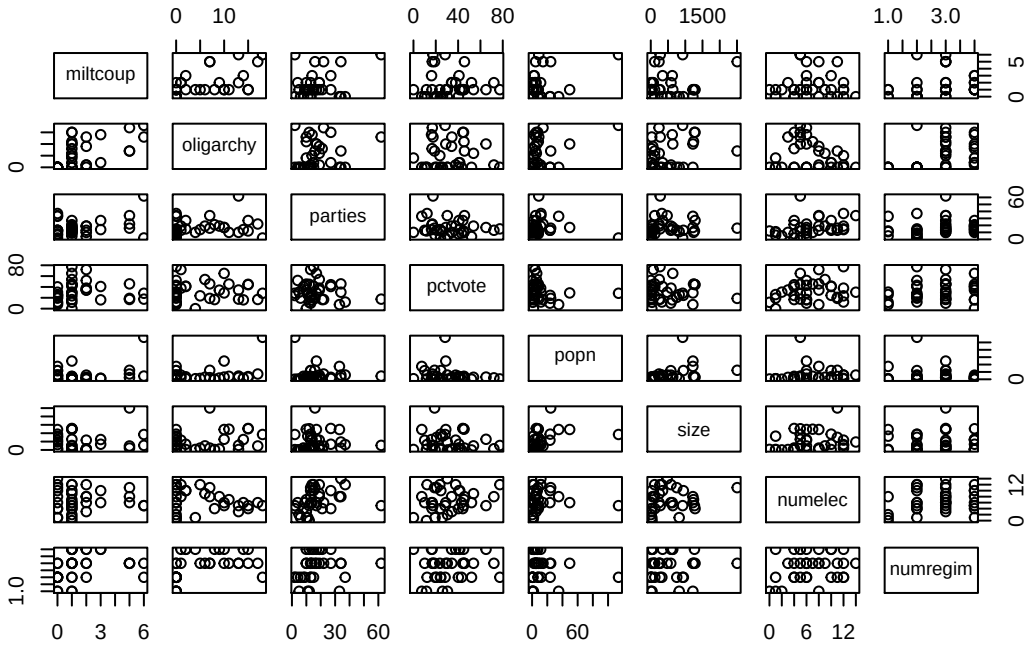
The first method is much more suitable for this kind of analyses, having the advantage that we can also see the influence that different factors have on the outcome. But of course that there are also disadvantages, mainly that we may take into account factors that do not have a big influence on our output (ex: Age).

Exercise 3.

a)



The distribution of the used variables, as well as the relation between independent variables were analysed and no indications for multicolinearity have been found. Some outliers for the variables parties, population and size were found. However, they were not removed due to the influence the removal might have on the sample size.




```
mean(miltcoup)
```

```
## [1] 1.583333
```

```
var(miltcoup)
```

```
## [1] 3.107143
```

A Poisson regression was performed on the full data set "africa", with the number of successful military coups ("miltcoup") as response variable. The other variables from the dataset are used as explanatory variables. Since the mean (=2.44) is similar to the variance (=2.13), the dependent variable seems to stem from a Poisson distribution.

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

```
##
## Call:
## glm(formula = miltcoup ~ oligarchy + pollib + parties + pctvote +
##     popn + size + numelec + numregim, family = poisson)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5075  -0.9533  -0.3100   0.4859   1.6459
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
```

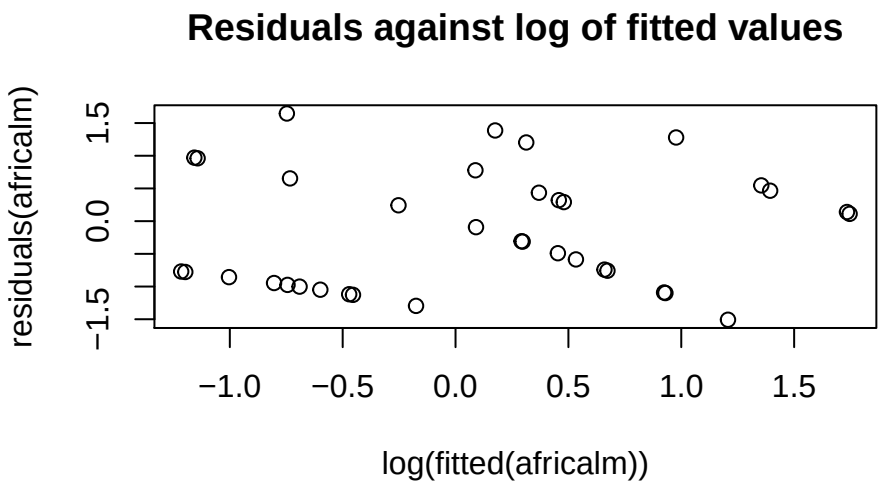
```
## (Intercept) -0.2334274  0.9976112  -0.234  0.81500
## oligarchy    0.0725658  0.0353457   2.053  0.04007 *
## pollib1     -1.1032439  0.6558114  -1.682  0.09252 .
## pollib2     -1.6903057  0.6766503  -2.498  0.01249 *
## parties      0.0312212  0.0111663   2.796  0.00517 **
## pctvote      0.0154413  0.0101027   1.528  0.12641
## popn         0.0109586  0.0071490   1.533  0.12531
## size        -0.0002651  0.0002690  -0.985  0.32444
## numelec     -0.0296185  0.0696248  -0.425  0.67054
## numregim     0.2109432  0.2339330   0.902  0.36720
## ---
## 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: 113.06
##
## Number of Fisher Scoring iterations: 5
```

```
pollib2 = as.numeric(pollib)
pollib2[pollib2 == 1] <- 4; pollib2[pollib2 == 3] <- 1; pollib2[pollib2 == 4] <- 3

pollib2 = as.factor(pollib2); contrasts(pollib2) = contr.sum
summary(glm(miltcoup~oligarchy+pollib2+parties+pctvote+popn+size+numelec+numregim, family=poisson))
```

```
##
## Call:
## glm(formula = miltcoup ~ oligarchy + pollib2 + parties + pctvote +
##     popn + size + numelec + numregim, family = poisson)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5075  -0.9533  -0.3100   0.4859   1.6459
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.1646106  0.8362809  -1.393  0.16374
## oligarchy    0.0725658  0.0353457   2.053  0.04007 *
## pollib21    -0.7591225  0.2822721  -2.689  0.00716 **
## pollib22    -0.1720607  0.2653712  -0.648  0.51674
## parties      0.0312212  0.0111663   2.796  0.00517 **
## pctvote      0.0154413  0.0101027   1.528  0.12641
## popn         0.0109586  0.0071490   1.533  0.12531
## size        -0.0002651  0.0002690  -0.985  0.32444
## numelec     -0.0296185  0.0696248  -0.425  0.67054
## numregim     0.2109432  0.2339330   0.902  0.36720
## ---
## 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: 113.06
##
## Number of Fisher Scoring iterations: 5
```



By analysing the results of the Poisson regression, it can be concluded that the factors "oligarchy", "parties" and "pollib" had a significant effect on the number of successful military coups from independence to 1989 and both oligarchy and parties had a positive impact on "miltcoup", as follows:

- the number of years the country was ruled by a military oligarchy (oligarchy): z = 2.053, p = 0.040
- the number of legal political parties ("parties"): z = 2.796, p = 0.005
- the political liberalization ("pollib")

Furthermore, it can be observed that with a political liberalization of full civil rights (2), the estimated number of successful coups is lower than for other political liberalizations, with z = -2.689 and p = 0.007.

b)

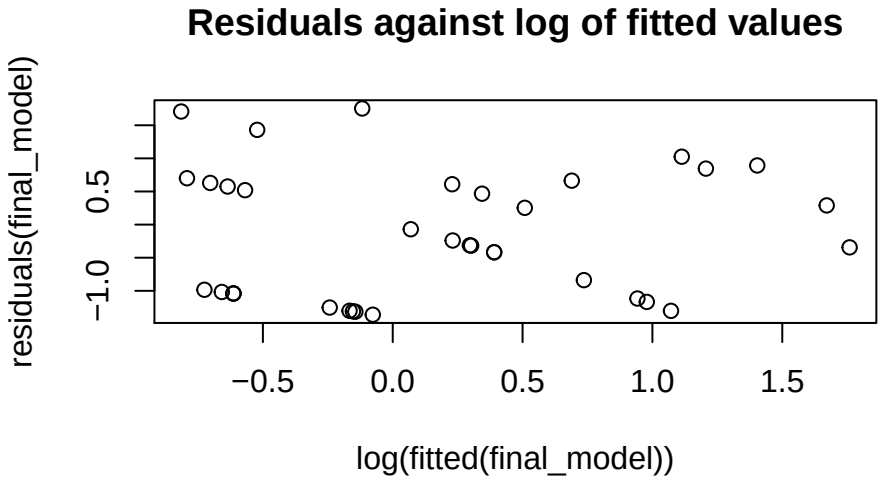
To reduce the number of explanatory variables, as most of the 8 independent variables are not significant for the output, the step down approach was used.

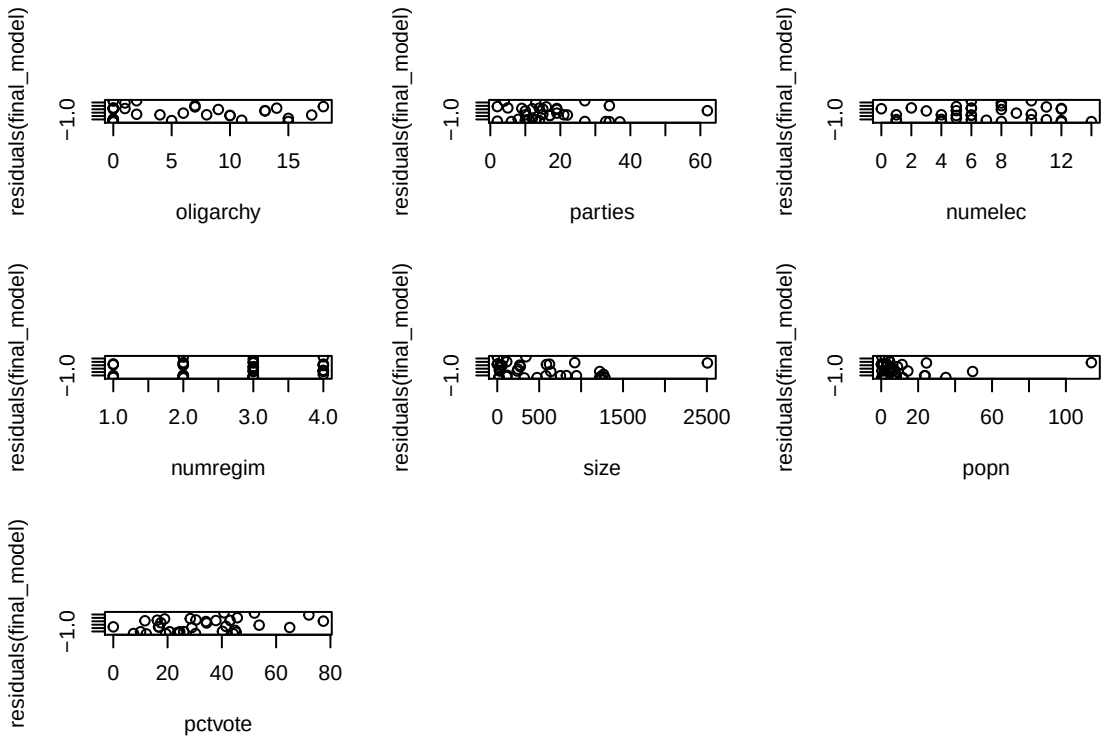
First, the model with all predictors was analysed. The variable "numelec" was removed first, as it it the the least significant. By repeating the process several times, "numregim", "size", "popn", "pctvote" have also been removed.

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

```
##
## Call:
## glm(formula = miltcoup ~ oligarchy + pollib + parties + pctvote +
##     popn + size + numelec + numregim, family = poisson, data = africa)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5075  -0.9533  -0.3100   0.4859   1.6459
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -0.2334274  0.9976112  -0.234  0.81500
## oligarchy    0.0725658  0.0353457   2.053  0.04007 *
## pollib1     -1.1032439  0.6558114  -1.682  0.09252 .
## pollib2     -1.6903057  0.6766503  -2.498  0.01249 *
## parties      0.0312212  0.0111663   2.796  0.00517 **
## pctvote      0.0154413  0.0101027   1.528  0.12641
## popn         0.0109586  0.0071490   1.533  0.12531
## size        -0.0002651  0.0002690  -0.985  0.32444
## numelec     -0.0296185  0.0696248  -0.425  0.67054
## numregim     0.2109432  0.2339330   0.902  0.36720
## ---
## 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: 113.06
##
## Number of Fisher Scoring iterations: 5
```





The resulting model consists of "oligarchy", "pollib" and "parties" - which were the only significant predictors in the complete model (from point a) as well, while the relation with the dependent variable remains the same as previously discussed. To investigate the distrubution of the residuals, they were plotted against the logarithm of the fitted values, yet no pattern was found in the obtained plot. To verify if any patterns emerge from the predictors, the residuals of the model were plotted against all the explanatory variables. However, the obtained plots showed no pattern that might indicate the need for a transformation (for the included variables) or an inclusion (for the excluded variables).