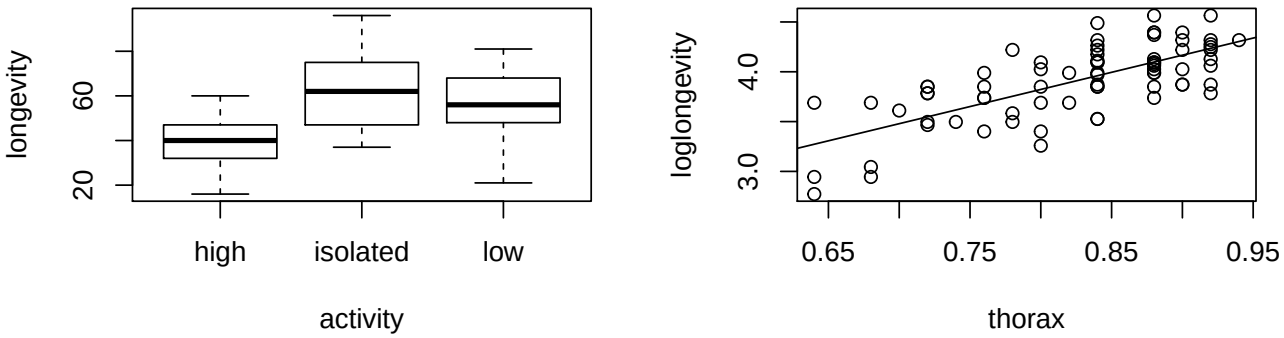
EDDA Group 29 Assignment 3

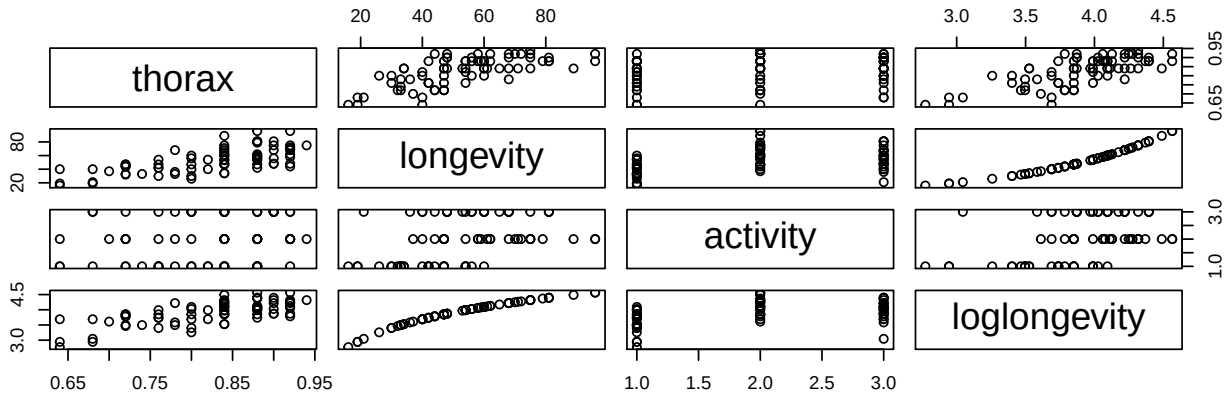
Geoffrey van Driessel (12965065), Yizhen Zhao (2658811) & Sophie Vos (2551583)

An overview of the R code is shown in the Appendix on page X.

Exercise 1

a) First, we add a column 'loglongevity' which will be used as a response variable. Next, we plot the longevity data in a separate boxplot for each activity. We observe that the longevity for fruitflies of the activity 'isolated' is the longest, followed by the activity 'low', the activity 'high' has the lowest longevity. Looking at the scater plot of loglongevity and thorax, we observe a weak linear correlation. The points follow a linear pattern, however, they are relatively widely spread. Furthermore, we could observe a weak linear correlation bewteen longevity and thorax.





In order to investigate whether sexual activity influences longevity, we performed an one-way Anova test. The null hypothesis states that sexual activity does not influence the longevity. The test results in a p-value smaller than the significance level of 0.05. Therefore, we reject H_0 and thus conclude that the sexual activity will influence the longevity. According to the summary, the estimated longevity for group 'high' is 3.60, and for group 'isolated' 3.60

+ 0.52 = 4.12 and for group 'low' 3.60 + 0.39 = 3.99. With a 95% confidence interval, the logevity for 'high' is [3.48 3.72], for 'isolated' [3.82, 4.41] and for 'low' [3.70, 4.29]. From this, we confirm that a high sexual acitivty has a negative impact on the longevity.

```
##
## Call:
## lm(formula = loglongevity ~ activity, data = fliesdata)
##
## 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
```

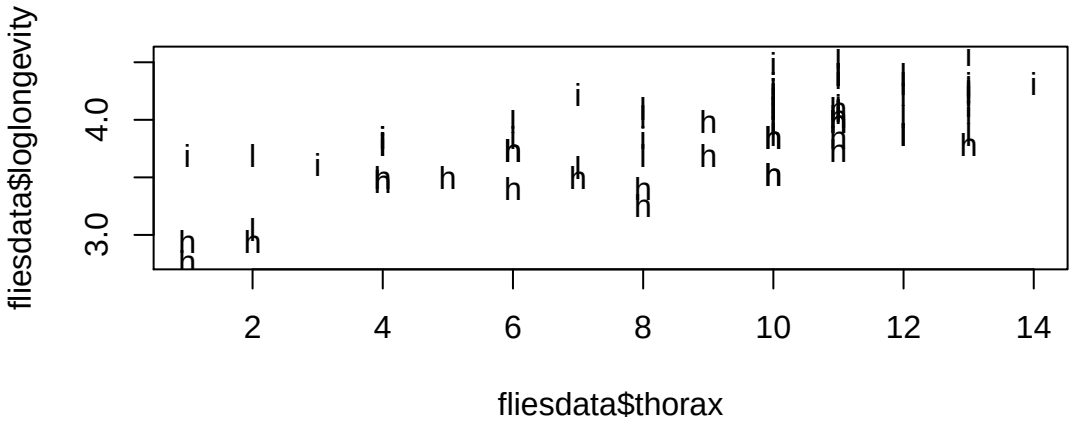
```
##                       2.5 %    97.5 %
## (Intercept)      3.4796296 3.7246190
## activityisolated 0.3439909 0.6904582
## activitylow      0.2244780 0.5709453
```

b) For this exercise, we apply two-way Anova with the two factors: activity and thorax. With H_0 (1) activity does not influence longevity, (2) thorax does not influence logevity, and (3) there is no interaction between activity and thorax. The output of this test shows that the p-values for the first two null hypotheses are all smaller than 0.05, therefore, we reject the first two null hypotheses. This means that activity and thorax influence the longevity. The p-value for the third null hypothesis is 0.4574. This is larger than 0.05, therefore, we do not reject the third null hypothesis. This means that there is no interaction betweem them. From the output we learned that the p-values for both activity and thorax are smaller than the significance level of 0.05. Therefore, H_0 is rejected which means that acitivity and thorax will effect the longevity. We calculated that the mean of thorax is equal to 0.82 and from the summary we observe that the estimated thorax is 2.98. Therefore, the estimated longevities for the three groups are: 'high' = (0.82 * 2.98) + 1.22 = 3.66, 'isolated' = (0.82 * 2.98) + 1.22 + 0.41 = 4.07 and, 'low' = (0.82 * 2.98) + 1.22 + 0.29 = 3.95. According to this result, we conclude that the higher activity is, the shorter longevity they have. This result is similar to the output in a).

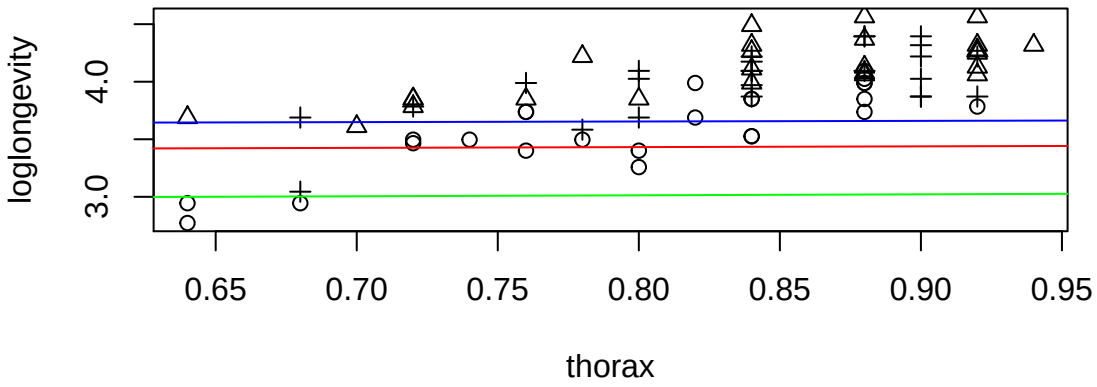
```
##
## Call:
## lm(formula = loglongevity ~ thorax + activity, data = fliesdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45369 -0.16746  0.02622  0.15306  0.33443
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.076233   0.067582  45.519  < 2e-16 ***
## thorax            0.067422   0.006934   9.724 1.10e-14 ***
## activityisolated 0.412046   0.058321   7.065 8.92e-10 ***
## activitylow       0.287140   0.058427   4.915 5.52e-06 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2026 on 71 degrees of freedom
## Multiple R-squared:  0.7214, Adjusted R-squared:  0.7096
## F-statistic: 61.29 on 3 and 71 DF,  p-value: < 2.2e-16
```

c) From the graph below we observe that longevity increases with the thorax. The group 'isolated' has the longest longevity, followed by 'low' and, lastly, 'high'.



Because thorax will influence the longevity, its dependence on activity is not so clear. Here we apply ANCOVA and use 'drop1' to obtain the p-value. According to the result, all the p-values are smaller than the significance level of 0.05. This confirms our analysis before, that both activity and thorax will influence the longevity. From the plot and summary below, we observe that the p-values for 'isolated:thorax' and 'low:thorax' are larger than the significance level of 0.05, therefore, we do not reject H_0. This means that there is no difference on thorax's dependence for the three activities. In other words, the dependence is similar under all three conditions of sexual activity.

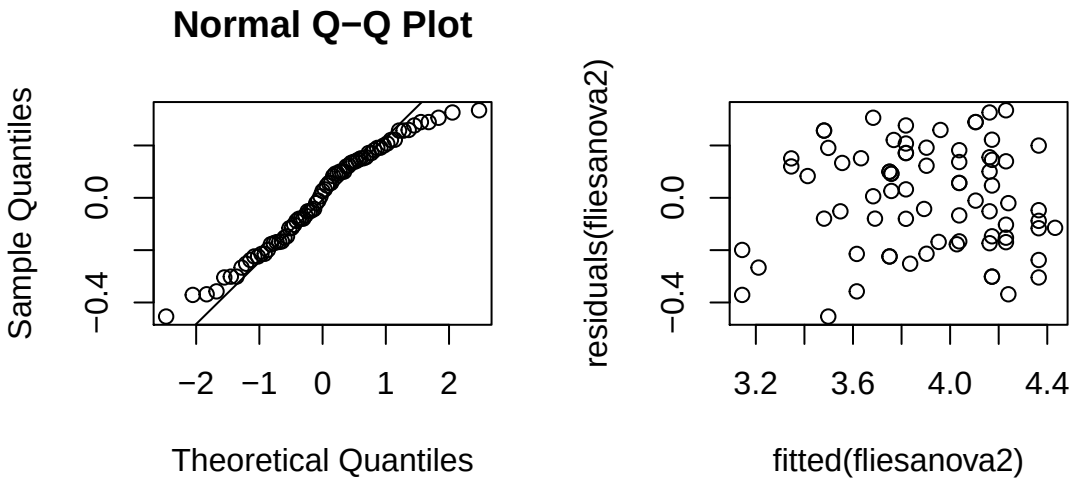



```
##
```

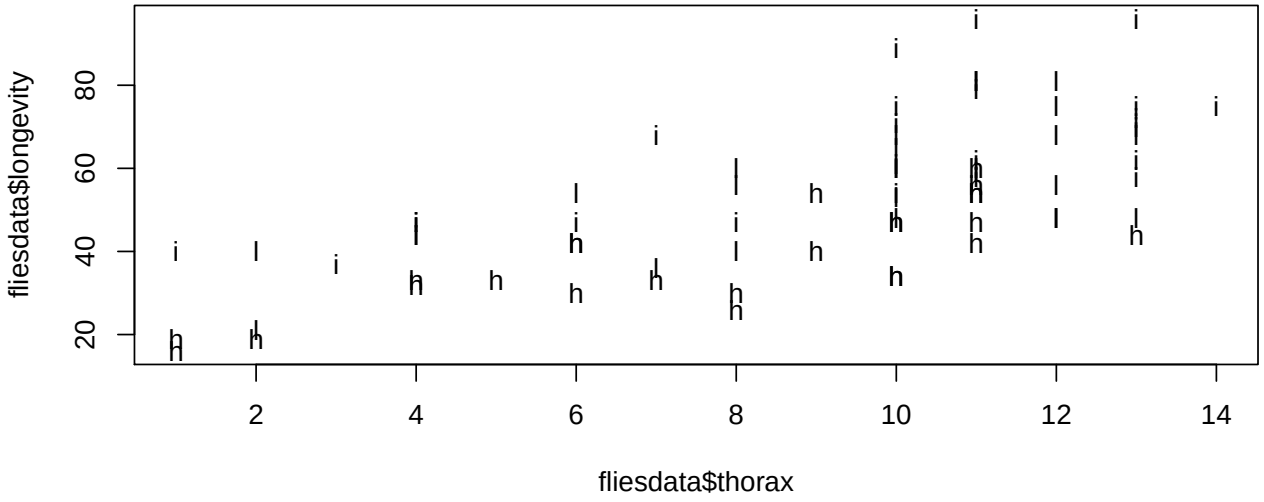
```
## Call:
## lm(formula = loglongevity ~ activity * thorax, data = fliesdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46624 -0.15549 -0.00804  0.15749  0.35592
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.312010   0.065954  50.217  < 2e-16 ***
## activity1         -0.366239   0.088099  -4.157 9.11e-05 ***
## activity2          0.298952   0.091817   3.256  0.00175 **
## thorax             0.068066   0.006929   9.823 9.69e-15 ***
## activity1:thorax   0.016082   0.009760   1.648  0.10397
## activity2:thorax  -0.013751   0.009405  -1.462  0.14827
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2006 on 69 degrees of freedom
## Multiple R-squared:  0.7345, Adjusted R-squared:  0.7153
## F-statistic: 38.18 on 5 and 69 DF,  p-value: < 2.2e-16
```

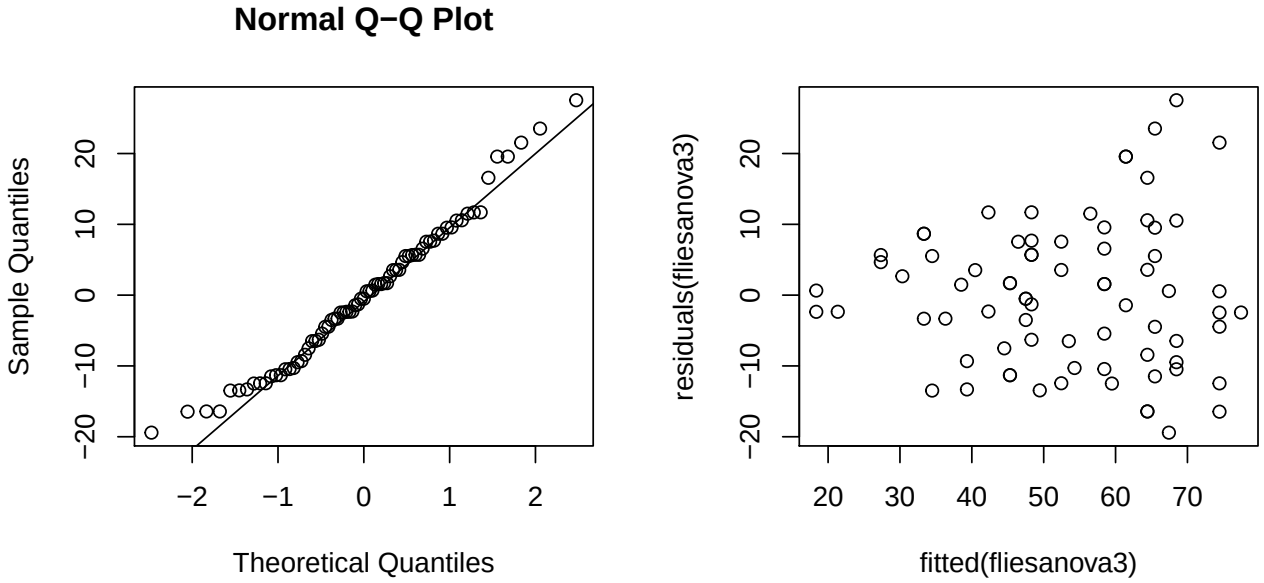
d) We prefer to take the length of the thorax into account. Based on our analysis above, we know that the length of the thorax influences the longevity of fruitflies. So it is not wise to ignore such a factor when doing analysis. But the first analysis is not wrong. At the begining, we did not know thorax's effect towards longevity and we only took one factor (activity) into account. Therefore, we apply one-way anova. Each test gives right results. As the first one only focus on activities' influence to longevity and second one focus on both activity and thorax.

e) Looking at the QQ-plot, we conclude that the data aprroximates a normal distribution. Looking at the residuals versus fitted values plot, we observe that there is no clear pattern. Therefore, we conclude that there is no sign of heteroscedasticity.



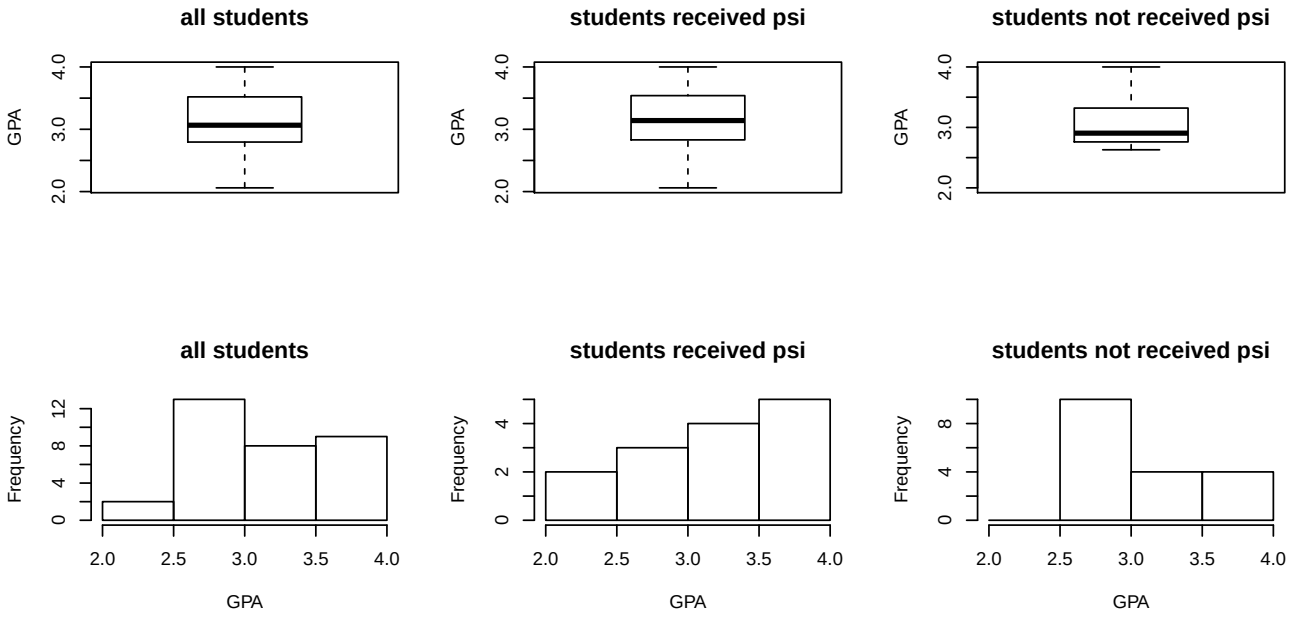
f) We do the same ancova analysis but use longevity as response variable. From the result we could know p-values for thorax and activity are smaller than significance level 0.05 therefore we get same conclusion as before that thorax and activity will effect fruitflies' longevity. Also we could see from the first plot that longevity increase with thorax. Then from the qq plot we could see the normality is also good. And from residuals versus fitted plot, we noticed some pattern and residuals seem to be bigger with bigger fitted values. So the inference here is, heteroscedasticity exists. In conclusion, it is wise to use the logarithm as response as we don't see heteroscedasticity in that model.



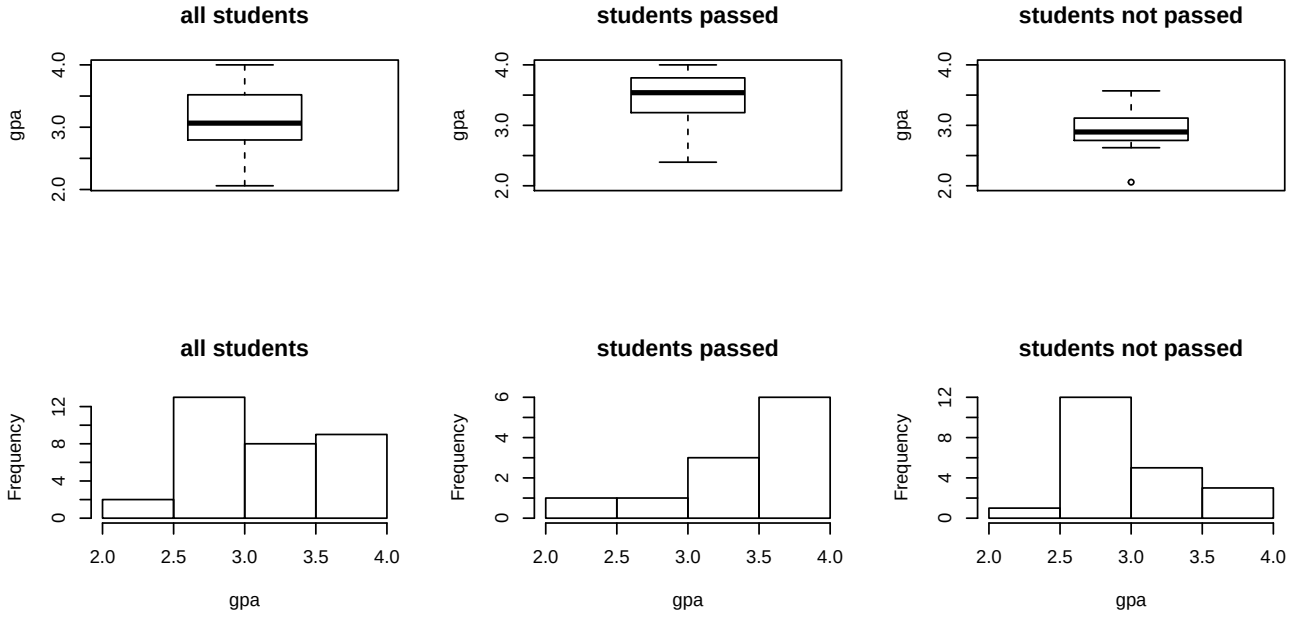


Exercise 2

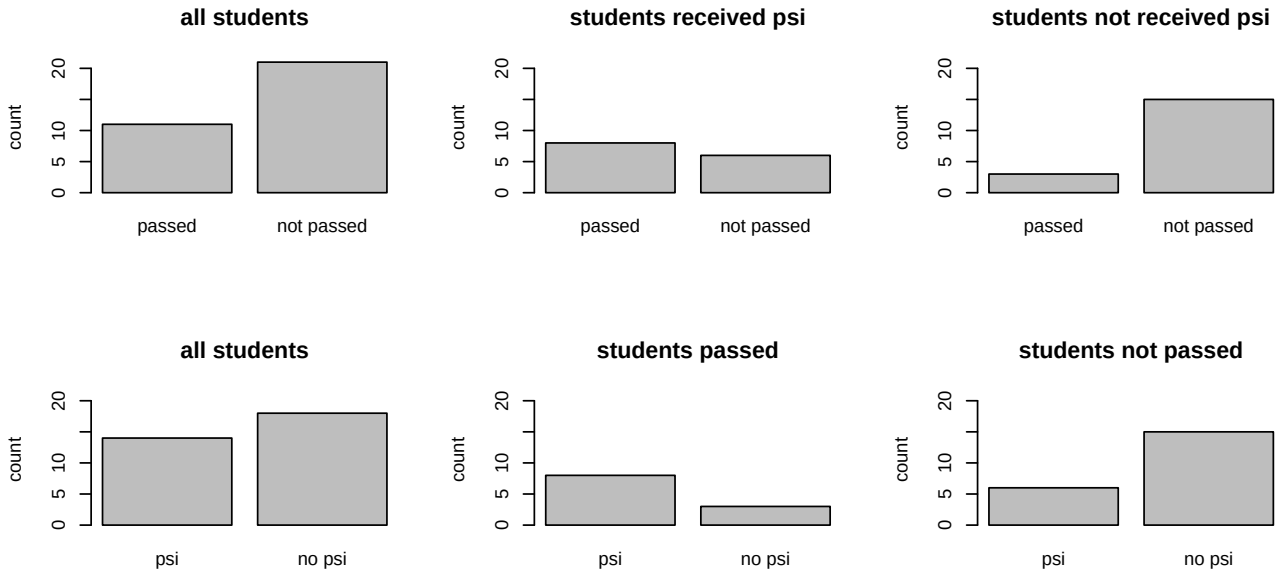
a) We study the data by exploring all combinations of the variables. First, we investigate the relation between the variables psi and gpa. We are interested whether the students that receive psi have a similar GPA to the students not receiving psi. We visualized the data in the boxplots below. We observe that the GPAs of all students is evenly distributed. The same applies to the GPAs of the students who received psi, however, the boxplot is positioned slightly higher. Looking at the boxplot of the students who did not receive psi, we observe that student with GPAs below 2.5 are not represented. Moreover, the boxplot is positioned lower compared to the others. To investigate the data further, we constructed histograms. We observe that for students who receive psi, the GPAs higher than 3.0 occur more frequently. In contrast, for students that did not receive psi, the GPAs between 2.5 and 3.0 occur more frequently. Hence, it can be argued that the data is biased because for the group of students who receive psi, the higher GPAs occur more frequently, whereas, for the group of students who do not receive psi, the lower GPAs occur more frequently.



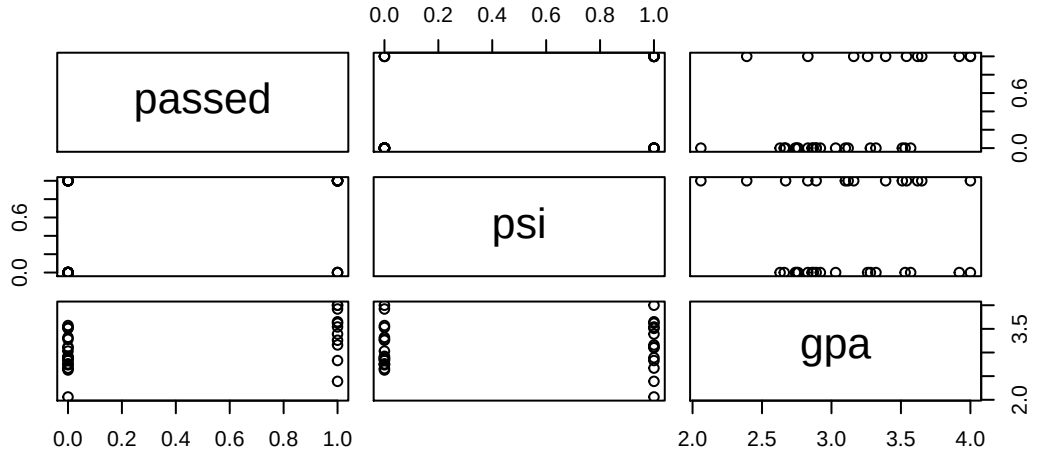
Next, we investigate the relation between the variables passed and gpa. Looking at the boxplots, we clearly see that students who passed the test have higher GPAs and the students who did not pass the test have lower GPAs. The histogram confirms this by showing higher frequencies of higher GPAs for students that passed the test and higher frequencies of lower GPAs for students that did not pass the test. Hence, it could be argued that students who have a higher GPA are more likely to pass the test.



Afterwards, we investigate the relation between the variables psi and passed. Looking at the bar plots below, we observe that more students did not pass the test compared to students who did pass the tests. In contrast, looking at the students who received psi, there are more students who passed than not passed, however, this difference is very small. For the students that did not receive psi, this difference is much larger and much more students did not pass compared to the students who passed. When considering all the students again, we observe that the amount of students receiving and not receiving psi is evenly distributed. Slightly more students did not receive psi compared to the students who received psi. Moreover, we observe that of the students who passed, more received psi and of the students who did not pass, more did not receive psi.



Lastly, we check the collinearity between all variables, and especially the explanatory factors. However, it is quite hard to spot collinearity with binominal data. For gpa we can see some sort of a positive relation with passed, and possibly a negative relation with psi. However, this visual diagnostics is too informal to conclude anything.



b) We fit a logistic regression model that explains if a student passes the test based on whether the student received psi and their gpa. We test the null hypotheses that receiving psi does not influence passing the assignment. According to the summary below, we observe that the p-value for psi is smaller than the significance level of 0.05. Therefore, we reject H_0. This means that psi works and does influences whether a student passes the test or not.

```
## Single term deletions
##
## Model:
## passed ~ gpa + psi
##        Df Deviance    AIC    LRT Pr(>Chi)
## <none>      26.253 32.253
## gpa     1   35.342 39.342 9.0885 0.002572 **
## psi     1   32.418 36.418 6.1647 0.013033 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Call:
## glm(formula = passed ~ gpa + psi, family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8396  -0.6282  -0.3045   0.5629   2.0378
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -11.602      4.213  -2.754  0.00589 **
## gpa            3.063      1.223   2.505  0.01224 *
## psi1           2.338      1.041   2.246  0.02470 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 41.183  on 31  degrees of freedom
## Residual deviance: 26.253  on 29  degrees of freedom
## AIC: 32.253
##
## Number of Fisher Scoring iterations: 5
```

c) Based on the summary of the model in b), we calculated the probability that a student with a gpa equal to 3 who receives psi or not passed the assignment.
For students who received psi:

$$\frac{1}{1+e^{-(-11.602+2.338)+(3.063*3)}} = 0.481$$

For students who did not receive psi:

$$\frac{1}{1+e^{-(-11.602+3.063*3)}} = 0.082$$

In conclusion, the probability for students with a gpa equal to 3 who receives psi, the probability of passing the assignment is 48.1%. For students with a gpa equal to 3 who do not receive psi, the probability of passing the assignment is 8.2%.

d) From the summary of the model in b), we notice that the coefficient of psi is 2.338, which is positive, this means that raising psi by 1 increases the linear predictor by 2.338 and increases the odds of passing the assignment by a factor $e^{2.338}$ which is equal to 10.36. This number means that students who receive psi are 10.36 times more likely to pass the assignment than those who do not receive psi. This is not dependent on gpa as gpa and psi are independent of each other.

e) We test the null hypothesis p1 = p2. This means that the null hypotheses state that students who do not receive psi and students who receive psi show the same improvement. In the matrix, we put the numbers 3, 15, 8 and 6. These numbers mean the following: from the 18 students who do not receive psi, 3 show improvement. This means that 18 - 3 = 15 students do not show improvements. From the 14 students who receive psi, 8 show improvement. This means that 14 - 8 = 6 students do not show improvement. Running Fisher's test when comparing the two binomial proportions, results in the p-value of 0.0265. This is smaller than the significance level of 0.05 and, therefore, H_0 is rejected. Thus, we conclude that students receiving and not receiving psi do not show a similar improvement.

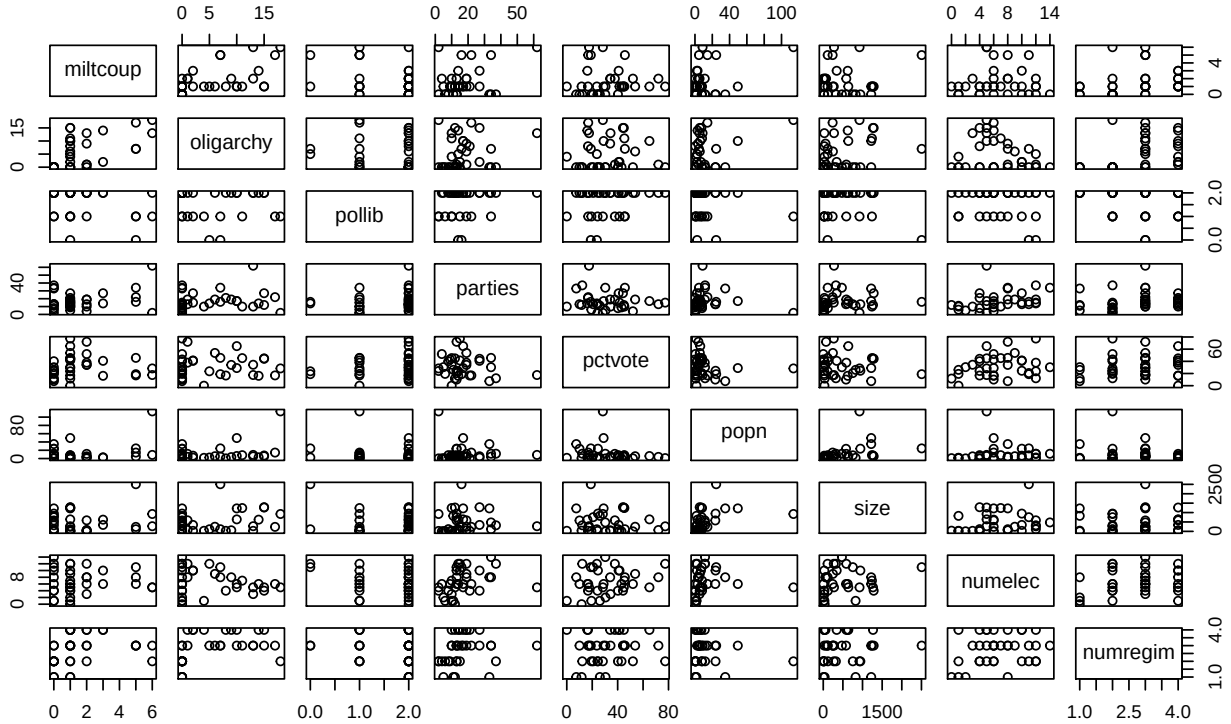
```
##
##  Fisher's Exact Test for Count Data
##
## data:  x
## p-value = 0.0265
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.02016297 0.95505763
## sample estimates:
## odds ratio
##  0.1605805
```

f) Yes, the second approach is not suitable as it ignores the influence of the gpa factor. With such a small dataset, the gpa could be heavily skewed/biased in one of the psi categories and the result of this would be that the chisquare test explains this bias with the difference in psi category, which is wrong. With a bigger dataset (central limit theorem: > 40) gpa will likely approximate a normal distribution and, therefore, we can assume that it will not have an influence. In this case, the chisquare test would be suitable.

g) An advantage of logistic regression is that it includes a predictive model which the Fisher exact test lacks. A disadvantage of logistic regression is that it needs all explanatory variables to be independent of each other. An advantage of Fisher's test is that it is a simpler test which is suitable with simpler datasets compared to logistic regression. A disadvantage of Fisher's test is that it can not make predictions and does not take other factors (blocks) into account.

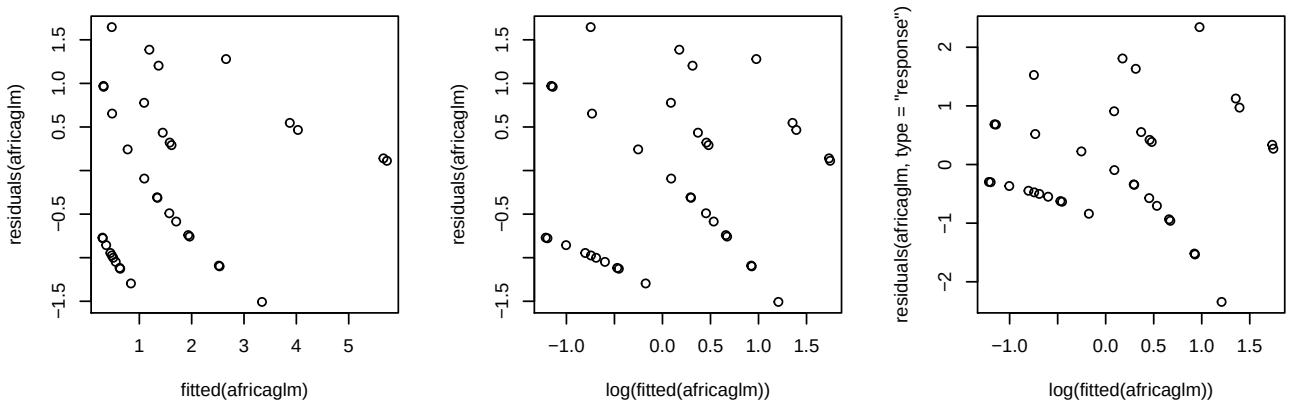
Exercise 3

a) First, we check if there are any linear correlated factors in the model by creating a scatterplot of all the variables. Looking at the scatterplot below, we conclude that there are no linear correlations. Afterwards, using the generalised linear regression model function, we run the Poisson regression. The output is presented below.

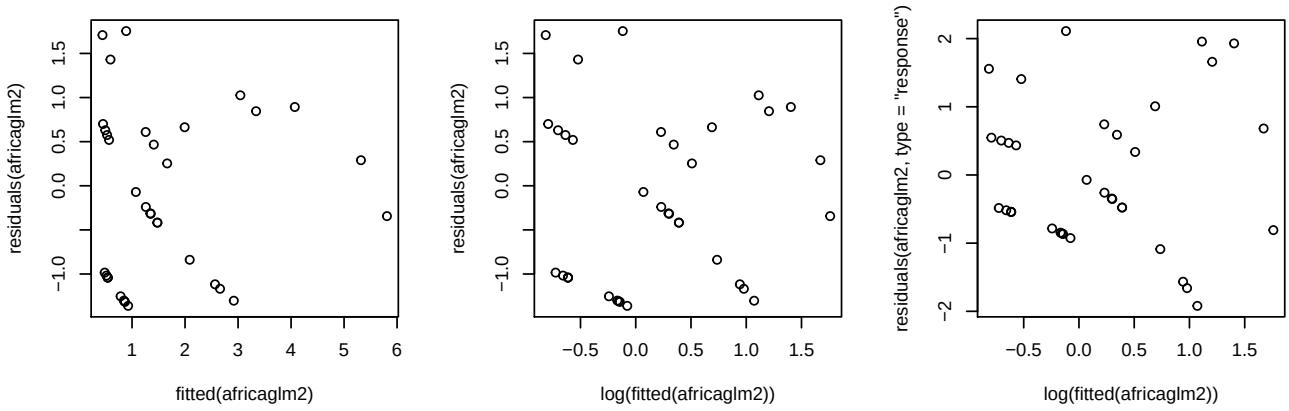


```
##
## Call:
## glm(formula = miltcoup ~ oligarchy + pollib + parties + pctvote +
##     popn + size + numelec + numregim, family = poisson, data = africa)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5075  -0.9533  -0.3100   0.4859   1.6459
##
## Coefficients:
##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

We conclude that oligarchy, pollib and parties significantly estimate (or have a linear relation with) the amount of successful military coups. As we take pollib as a factor, we find that category 2 (full civil rights) has significant less military coups (estimated 1.69 coups less) than pollib category 0. Afterwards, to evaluate the model, we plotted the residuals against the fitted values. The plot shows equal variance, however, a pattern can be observed. This is due to the dependent variable being a count on a small scale (0 - 6) which can be interpreted as discrete data. Approximately, for each target value, a curve is visible. Next, we calculate the logarithm to ensure that the x-values are fitted by a linear function. The second plot shows more spread, however, the previously mentioned structure of curves is still visible. Finally, we plot the response residuals. We observe that the response residuals increase with the (logarithm) of the fitted values, as expected under a Poisson model.



b) Following the step down method, we removed the factors in the order: numelec > numregim > size > popn > pctvote. This results in the model miltcoup = 0.251377 + 0.092622 * oligarchy - 0.574103 * pollib + 0.022059 * parties + error. In this process, we started with an R-squared value of 0.57 and ended up with a value of 0.50, however, we reduced the model from eight factors to three. Moreover, the residual plots look similar to the ones in a) in which all factors were included in the model.



Appendix: R code

```
# --- Exercise 1 --- #

# --- Exercise 2 --- #

#A
#psi vs gpa
data = read.table("psi.txt", header = TRUE);
data_psi = subset(data, psi == 1)
data_no_psi = subset(data, psi == 0)
par(mfrow=c(2,3))
boxplot(data$gpa,ylab="GPA",main="all students")
boxplot(data_psi$gpa,ylab="GPA",main="students received psi")
boxplot(data_no_psi$gpa,ylab="GPA",main="students not received psi",ylim=c(2.0,4.0))
hist(data$gpa,xlab="GPA",main="all students")
```

```r
hist(data_psi$gpa,xlab="GPA",main="students received psi")
hist(data_no_psi$gpa,xlab="GPA",main="students not received psi",
  breaks = c(2.0,2.5,3.0,3.5,4.0))
# passed vs gpa
par(mfrow=c(2,3))
data_passed = subset(data, passed == 1)
data_not_passed = subset(data, passed == 0)
boxplot(data$gpa,ylab="gpa",main="all students")
boxplot(data_passed$gpa,ylab="gpa",main="students passed",ylim=c(2.0,4.0))
boxplot(data_not_passed$gpa,ylab="gpa",main="students not passed",ylim=c(2.0,4.0))
hist(data$gpa,xlab="gpa",main="all students")
hist(data_passed$gpa,xlab="gpa",main="students passed")
hist(data_not_passed$gpa,xlab="gpa",main="students not passed", breaks = c(2.0,2.5,3.0,3.5,4.0))
# passed vs psi
par(mfrow=c(2,3))
barplot(c(nrow(data_passed),nrow(data_not_passed)),ylim=c(0,21),main="all students",
  ylab="count",names.arg=c("passed","not passed"))
barplot(c( nrow(subset(data_psi, passed == 1)),nrow(subset(data_psi,passed == 0))),
  ylim=c(0,21),main="students received psi",ylab="count",names.arg=c("passed","not passed"))
barplot(c( nrow(subset(data_no_psi, passed == 1)),nrow(subset(data_no_psi,passed == 0))),
  ylim=c(0,21),main="students not received psi",ylab="count",names.arg=c("passed","not passed"))
barplot(c(nrow(data_psi),nrow(data_no_psi)),ylim=c(0,21),main="all students",
  ylab="count",names.arg=c("psi","no psi"))
barplot(c( nrow(subset(data_passed, psi == 1)),nrow(subset(data_passed,psi == 0))),
  ylim=c(0,21),main="students passed",ylab="count",names.arg=c("psi","no psi"))
barplot(c( nrow(subset(data_not_passed, psi == 1)),nrow(subset(data_not_passed,psi == 0))),
  ylim=c(0,21),main="students not passed",ylab="count",names.arg=c("psi","no psi"))
#B
data$passed = factor(data$passed)
data$psi = factor(data$psi)
model <- glm(passed~gpa+psi,data=data,family=binomial)
drop1(model,test="Chisq")
summary(model)
#E
x=matrix(c(3,15,8,6),2,2)
fisher.test(x)


# --- Exercise 3 --- #

#A
africa = read.table("africa.txt", header = TRUE)
plot(africa)
africa$pollib = factor(africa$pollib)
africaglm=glm(miltcoup~oligarchy+pollib+parties+pctvote+popn+size+numelec+numregim,
  family=poisson,data=africa)
summary(africaglm)
par(mfrow=c(1,3))
plot(fitted(africaglm),residuals(africaglm))
plot(log(fitted(africaglm)),residuals(africaglm))
plot(log(fitted(africaglm)),residuals(africaglm, type="response"))
#B
summary(glm(miltcoup~oligarchy+pollib+parties+pctvote+popn+size+numelec+numregim,
  family=poisson,data=africa))
summary(glm(miltcoup~oligarchy+pollib+parties+pctvote+popn+size+numregim,
  family=poisson,data=africa))
```

```r
summary(glm(miltcoup~oligarchy+pollib+parties+pctvote+popn+size,
  family=poisson,data=africa))
summary(glm(miltcoup~oligarchy+pollib+parties+pctvote+popn,
  family=poisson,data=africa))
summary(glm(miltcoup~oligarchy+pollib+parties+pctvote,
  family=poisson,data=africa))
summary(glm(miltcoup~oligarchy+pollib+parties,
  family=poisson,data=africa))
africaglm2=glm(miltcoup~oligarchy+pollib+parties,
  family=poisson,data=africa)
with(summary(africaglm2), 1 - deviance/null.deviance)
summary(africaglm2)
par(mfrow=c(1,3))
plot(fitted(africaglm2),residuals(africaglm2))
plot(log(fitted(africaglm2)),residuals(africaglm2))
plot(log(fitted(africaglm2)),residuals(africaglm2, type="response"))
```