Assignment 3

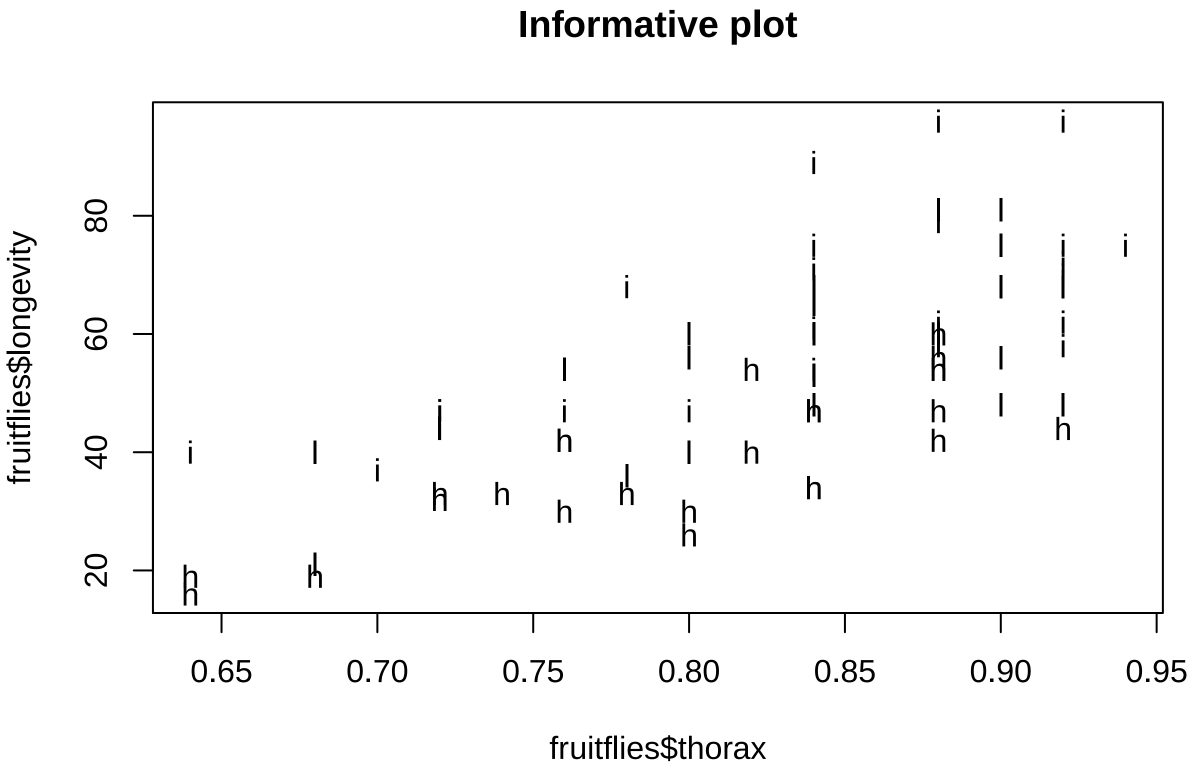
Andrea Di Dio, Chenghan Song, Jiacheng Lu — Group 22

Exercise 1

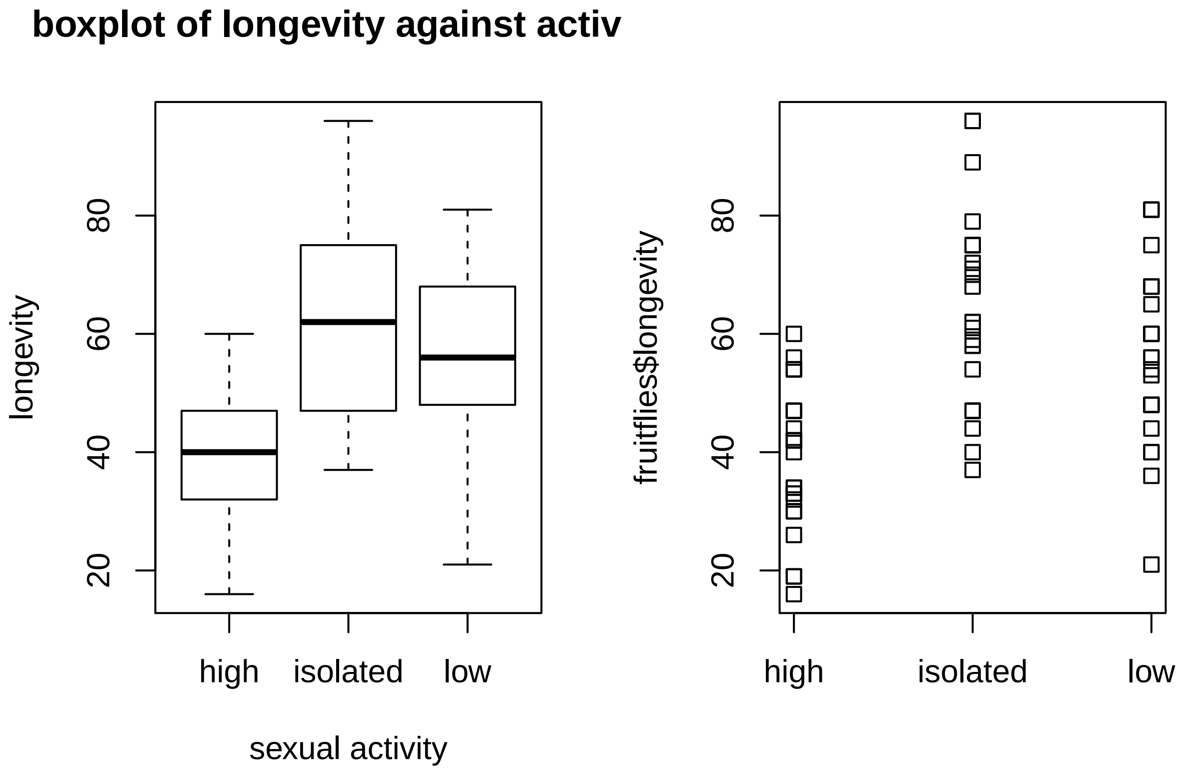
a)

First add loglongevity to the dataframe. We draw plots to show the graphical summary of the data.
H_0 : The sexual activity does not affect longevity.

```
data=read.table(file="fruitflies.txt",header=TRUE)
fruitflies=data.frame(data)
fruitflies$loglongevity=log(fruitflies$longevity,exp(1))
plot(fruitflies$longevity~fruitflies$thorax,pch=as.character(fruitflies$activity),main="Informative plo
```



```
par(mfrow=c(1,2))
boxplot(fruitflies$longevity~fruitflies$activity,data=fruitflies,xlab="sexual activity",ylab="longevity
stripchart(fruitflies$longevity~fruitflies$activity,vertical=TRUE)
```



```
fruitfliesaov=lm(loglongevity~activity,data=fruitflies)
anova(fruitfliesaov)
```

```
## 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
```

```
confint(fruitfliesaov)
```

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

The p-value for testing H_0 is smaller than 0.05, thus we reject H_0. Which means regardless of thorax, sexual activity influences longevity. Use confidence interval to estimate longevity. For high sexual activity, estimated longevity is between 32 and 41; for isolated activity, longevity is between 46 and 83; for low sexual activity, longevity is between 41 and 73.

b)

H_0:The sexual activity does not affect longevity.

```
fruitflies$activity=as.factor(fruitflies$activity)
fruitfliesaov2=lm(loglongevity~thorax+activity,data=fruitflies)
drop1(fruitfliesaov2,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
```

The p-value for testing H_0 is smaller than 0.05, thus we reject H_0. Which means Taking thorax into consideration, sexual activity still has an influence on longevity.

```
contrasts(fruitflies$activity)=contr.sum
fruitfliesaov3=lm(loglongevity~thorax+activity,data=fruitflies)
summary(fruitfliesaov3)
```

```
##
## Call:
## lm(formula = loglongevity ~ thorax + activity, data = fruitflies)
##
## 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)  1.45083    0.25392   5.714 2.41e-07 ***
## thorax       2.97899    0.30665   9.715 1.14e-14 ***
## activity1   -0.23189    0.03395  -6.831 2.40e-09 ***
## activity2    0.17809    0.03329   5.349 1.03e-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
```

```
mean(fruitflies$thorax)
```

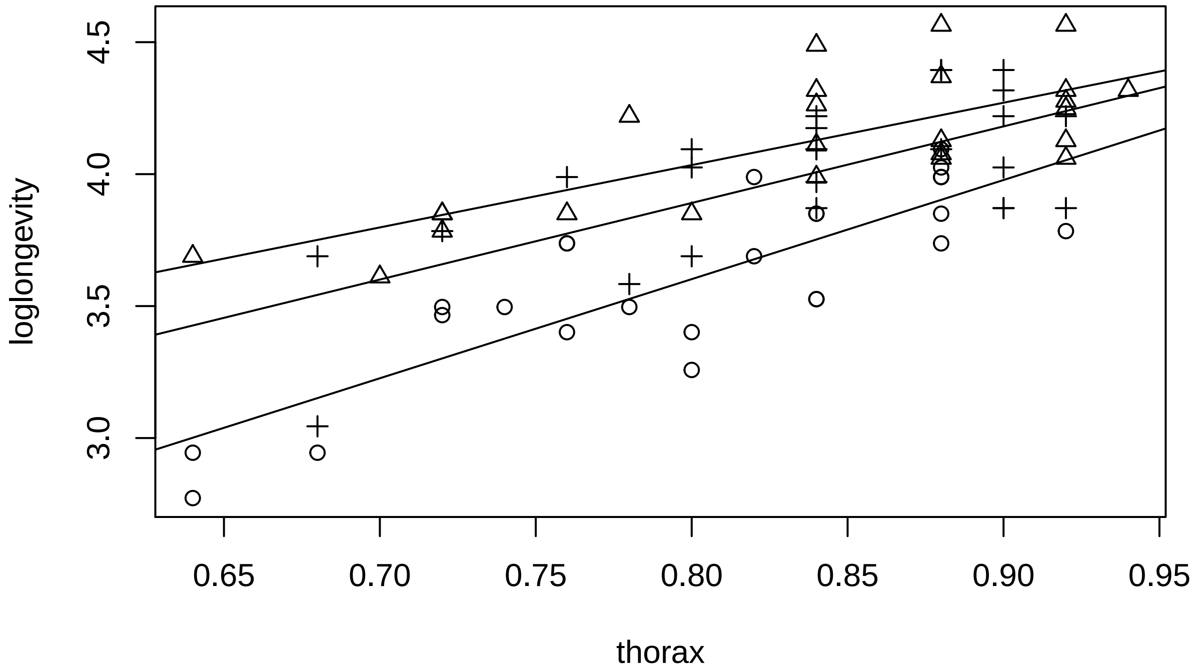
```
## [1] 0.8245333
```

We can calculate $\alpha_3 = -\alpha_1 - \alpha_2$. From the result of the test, we can find out that $alpha_1 < alpha_3 < alpha_2$, which means sexual activity decrease longevity. $\mu = 1.45083, \beta = 2.97899, \alpha_1 = -0.23189, \alpha_2 = 0.17809, average_thorax = 0.824533$. According to the model, for high sexual activity, the estimated longevity is about 39.46, for isolated sexual activity is about 59.47, for low sexual activity is about 52.50.

c)

H_0:The thorax does not affect longevity.

```
plot(loglongevity~thorax,pch=unclass(fruitflies$activity),data=fruitflies)
abline(lm(loglongevity~thorax,data=fruitflies[c(1:25),]))
abline(lm(loglongevity~thorax,data=fruitflies[c(26:50),]))
abline(lm(loglongevity~thorax,data=fruitflies[c(51:75),]))
```



```
aovfruitfliesinter=lm(loglongevity~activity*thorax,data=fruitflies)
summary(aovfruitfliesinter)
```

```
##
## Call:
## lm(formula = loglongevity ~ activity * thorax, data = fruitflies)
##
## 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)        1.4372     0.2535   5.669 3.08e-07 ***
## activity1         -0.8394     0.3505  -2.395   0.0194 *
## activity2          0.7071     0.3458   2.045   0.0447 *
## thorax             3.0065     0.3060   9.827 9.55e-15 ***
## activity1:thorax   0.7489     0.4293   1.744   0.0855 .
## activity2:thorax  -0.6440     0.4147  -1.553   0.1250
## ---
## 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
```

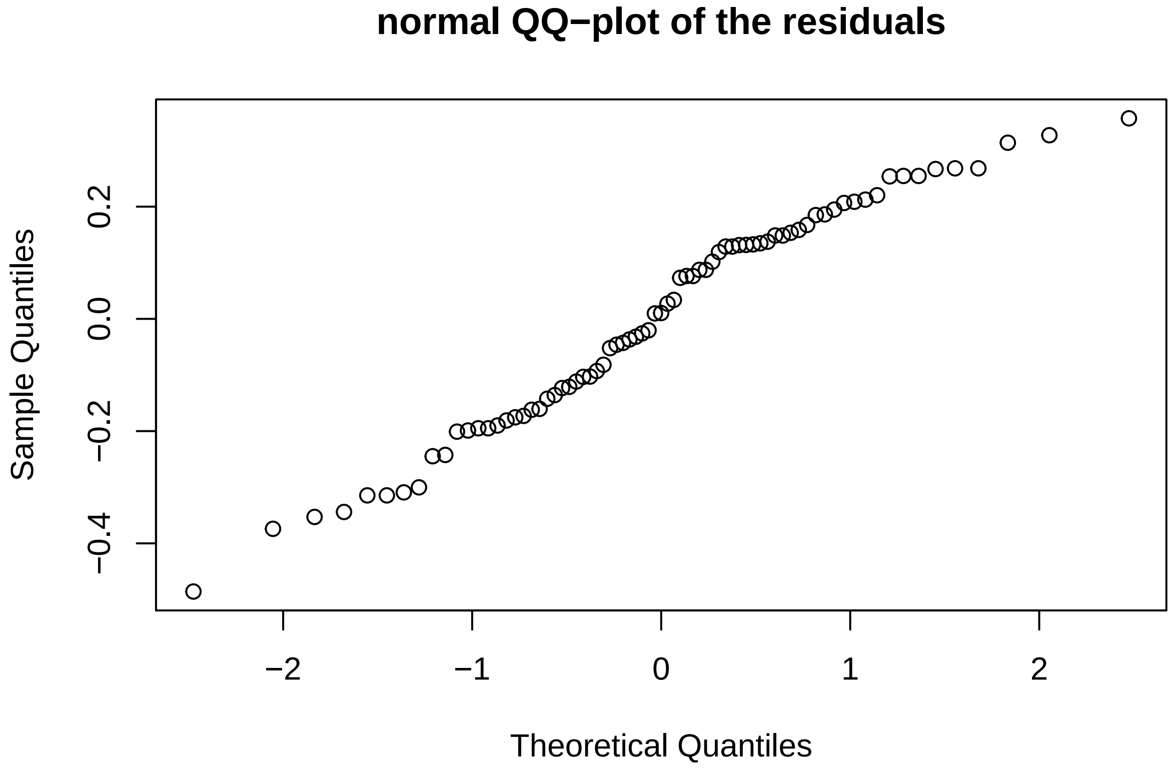
From the test result, the estimate of activity1:thorax and activity2:thorax give the estimated differences. In the plot we can see that the lines are not parallel. The dependence is not similar under three conditions of sexual activity.

d)

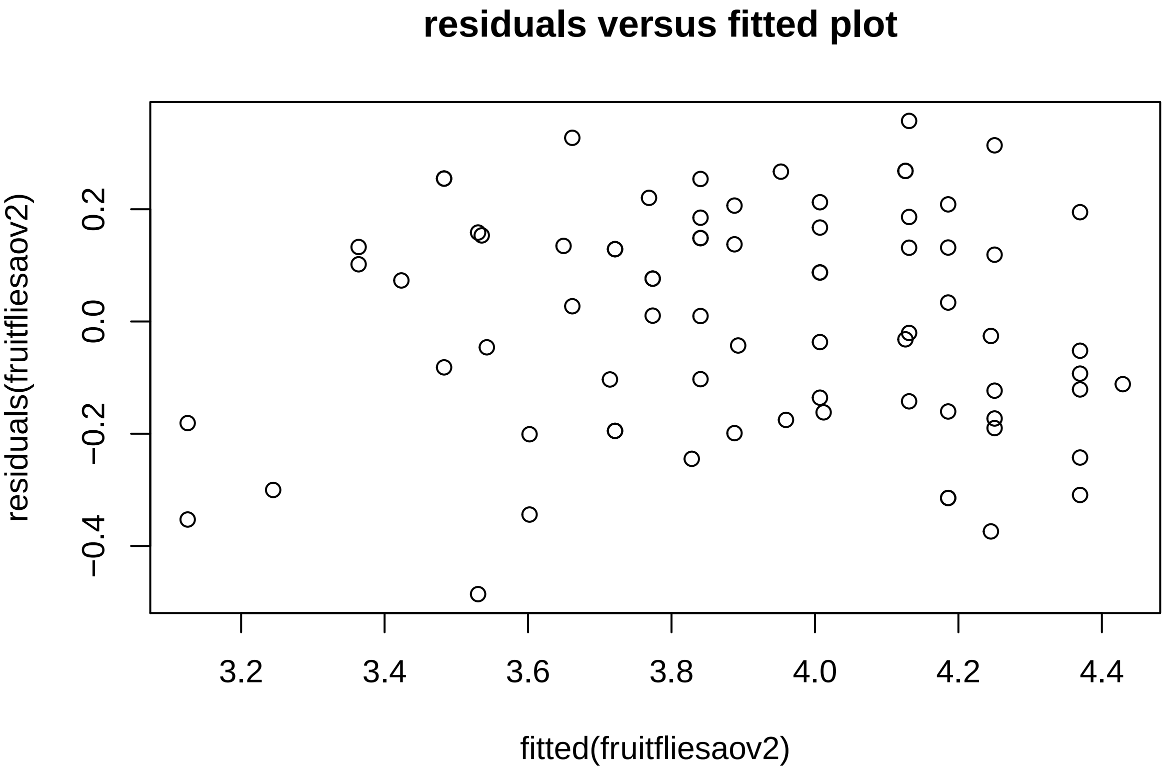
We think the analyses with thorax length is better because from the test result in c) we can see that thorax length does influence longevity, so we should take thorax length into consideration.

e)

```
qqnorm(residuals(fruitfliesaov2),main="normal QQ-plot of the residuals")
```



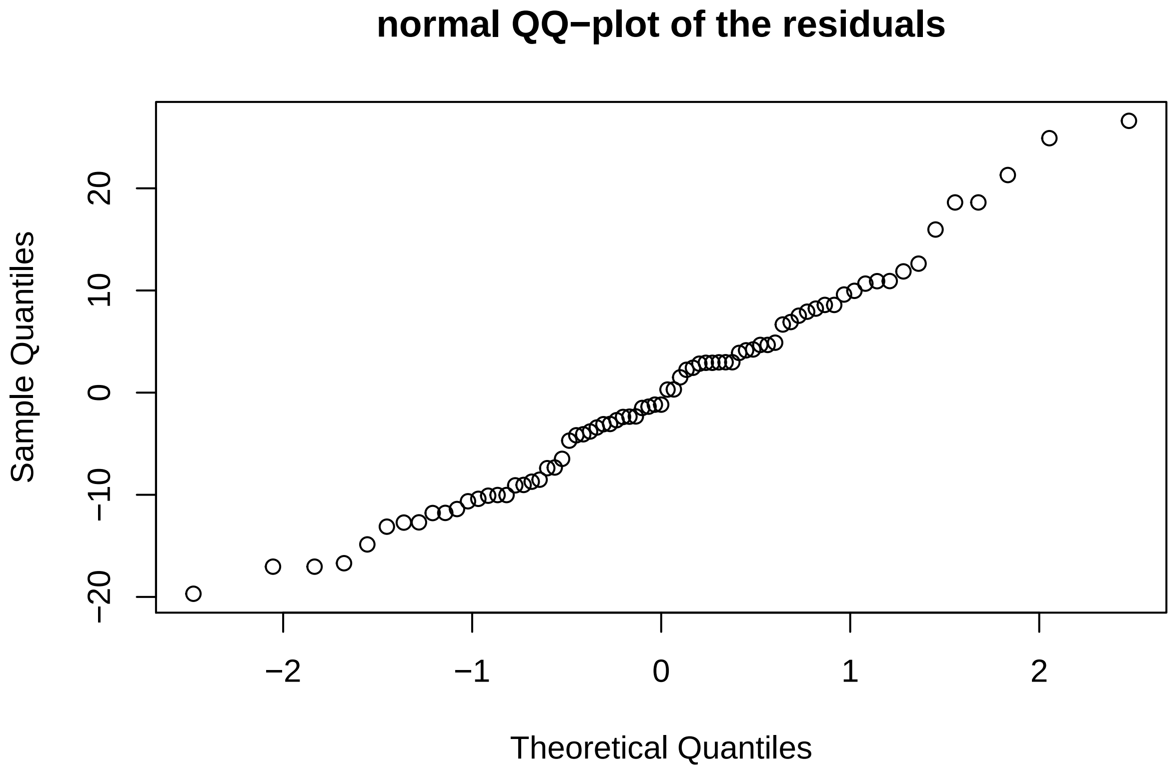
```
plot(fitted(fruitfliesaov2),residuals(fruitfliesaov2),main=" residuals versus fitted plot")
```



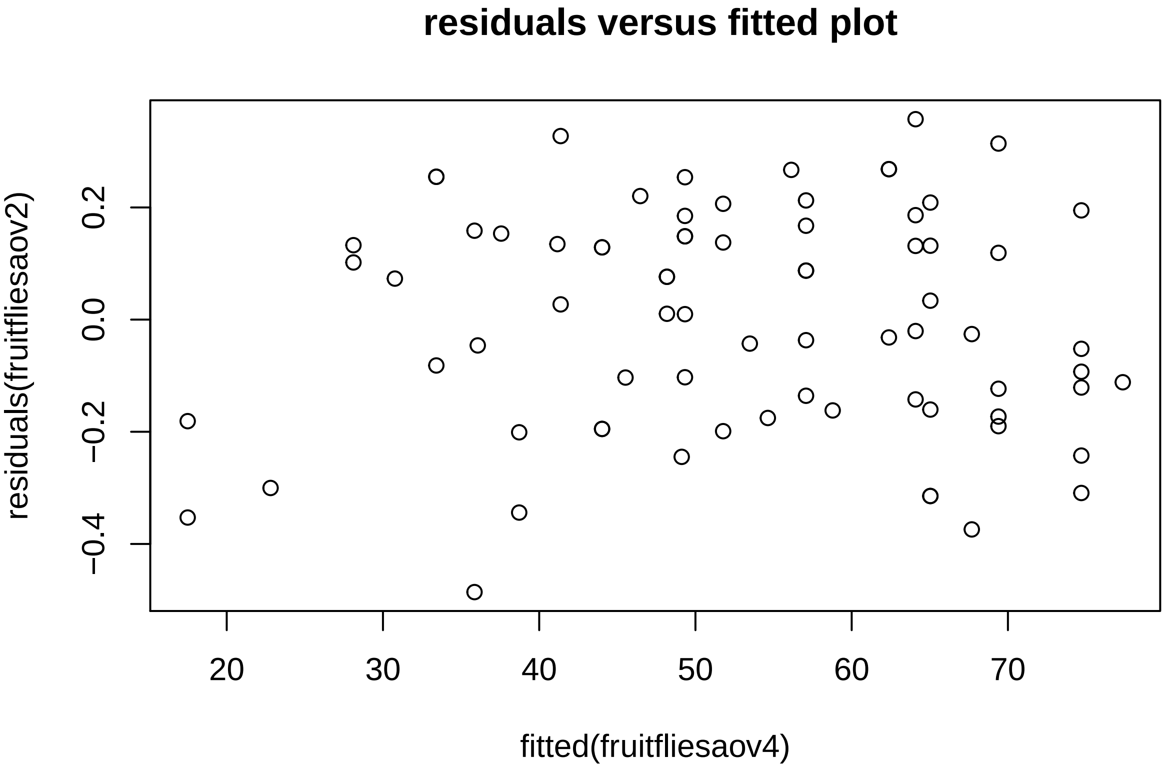
From the normal qq plot of the residuals we can see that the spread is not normal. The plot of residuals versus fitted plot does not show any pattern which means no heteroscedasticity.

f)

```
fruitflies$activity=as.factor(fruitflies$activity)
fruitfliesaov4=lm(longevity~thorax+activity,data=fruitflies)
qqnorm(residuals(fruitfliesaov4),main="normal QQ-plot of the residuals")
```



```
plot(fitted(fruitfliesaov4),residuals(fruitfliesaov2),main=" residuals versus fitted plot")
```

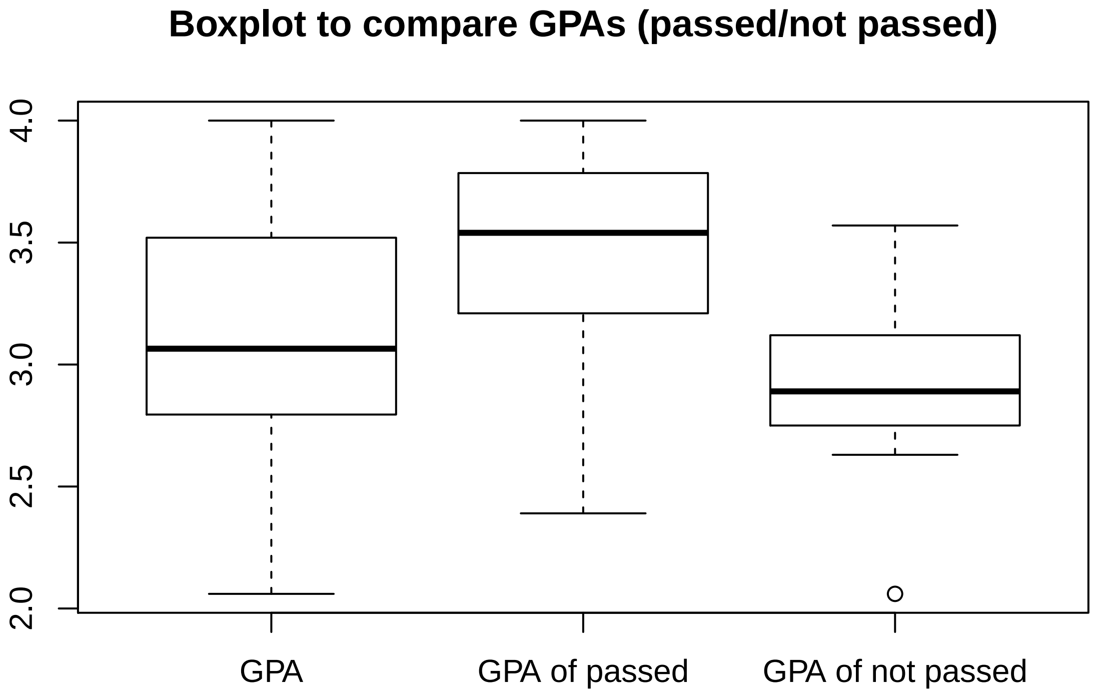


If we use longevity rather than its logarithm, from the qq plot we can see that it is normally distributed. Comparing with logarithm as response, it is clear that logarithm is much better.

Exercise 2

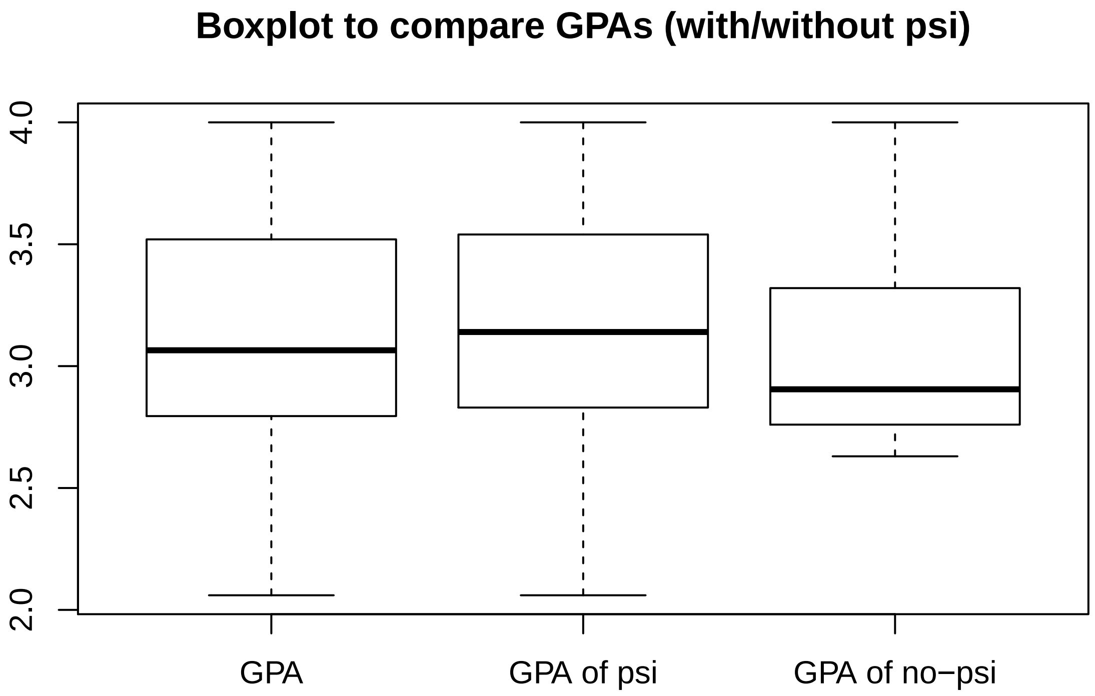
a)

```
psi_raw <- read.delim("./psi.txt", header = TRUE, sep = "")
psi_yes <- psi_raw[which(psi_raw$psi == 1),]
psi_no <- psi_raw[which(psi_raw$psi == 0),]
passed <- psi_raw[which(psi_raw$passed == 1),]
not_passed <- psi_raw[which(psi_raw$passed == 0),]
boxplot(psi_raw$gpa, passed$gpa, not_passed$gpa, names = c("GPA","GPA of passed","GPA of not passed"),
        main = "Boxplot to compare GPAs (passed/not passed)")
```



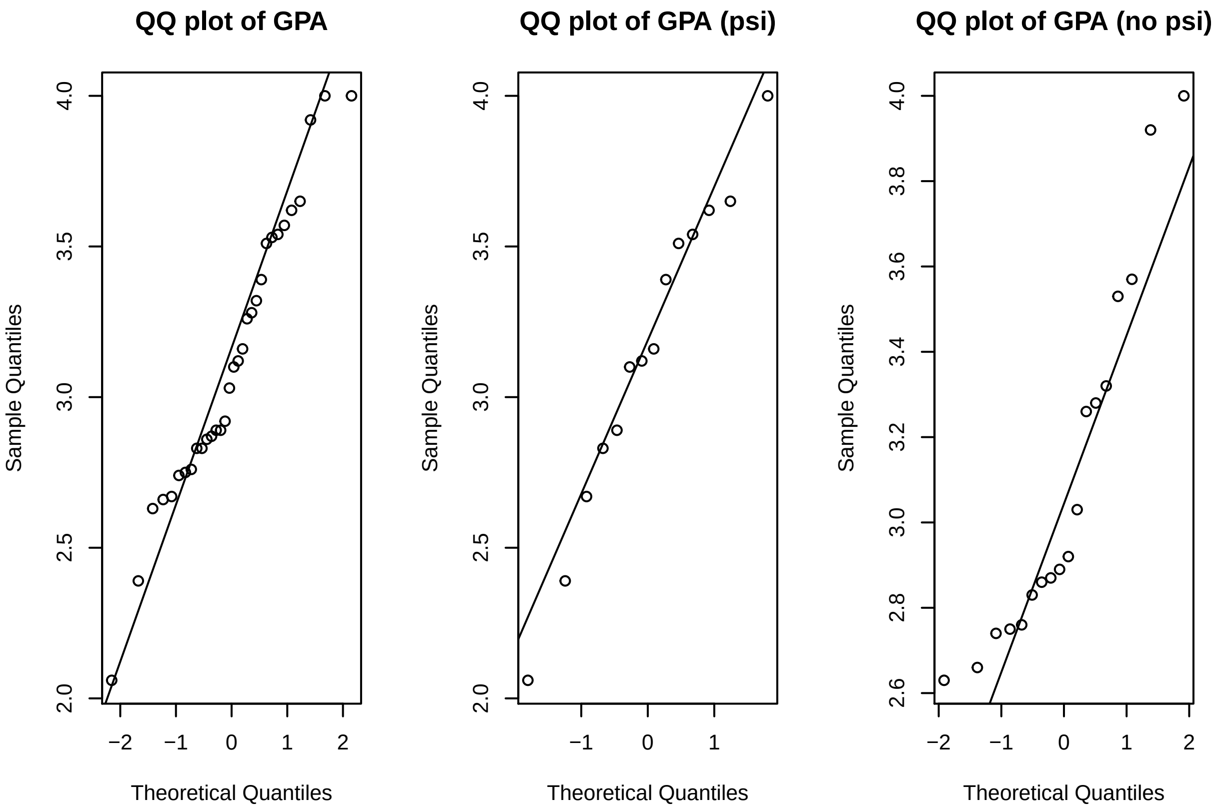
The boxplot above shows the comparison among the total GPA of the students, the GPA for the students who passed the assignments and the GPA for the students that did not pass the assignment. The boxplot shows that the median GPA of the students who passed the assignment is higher than that of the students who didn't pass regardless of the teaching method. This could mean that the GPA of a student is a significant factor to the outcome of an assignment for the students.

```
boxplot(psi_raw$gpa, psi_yes$gpa, psi_no$gpa, names = c("GPA", "GPA of psi", "GPA of no-psi"),
        main = "Boxplot to compare GPAs (with/without psi)")
```



The boxplot above shows the comparison among the total GPA of the students, the GPA for the students who were taught using *psi* and the GPA for the students who were taught using the traditional method. It shows that the median GPA of a student who has been taught with *psi* is marginally higher than that of the students who were taught using the traditional method. This could mean that the teaching method does not necessarily influence the GPA for a student.

```r
par(mfrow = c(1,3))
qqnorm(psi_raw$gpa, main = "QQ plot of GPA");qqline(psi_raw$gpa)
qqnorm(psi_yes$gpa, main = "QQ plot of GPA (psi)");qqline(psi_yes$gpa)
qqnorm(psi_no$gpa, main = "QQ plot of GPA (no psi)");qqline(psi_no$gpa)
```



The QQ-plots above compare the population distribution of the GPA among all students, the students who received *psi* and the students who were taught using the traditional method. The only QQ-plot which might indicate normality of the population is the one showing the GPA for the students who have been taught with the *psi* method whilst the other two plots don't seem to be sampled from a normal distribution. This might be a problem given that the sample population of the data for students taught with *psi* and those taught with the traditional method seems to be different making the experiment less fair or significant.

```
psi_passed_tab <- xtabs(~psi_raw$passed + psi_raw$psi)
rownames(psi_passed_tab) <- c("Passed", "Not Passed")
colnames(psi_passed_tab) <- c("No psi", "psi")
ftable(psi_passed_tab)
```

```
##                 psi_raw$psi No psi psi
## psi_raw$passed
## Passed                          15   6
## Not Passed                       3   8
```

The contingency table above shows that the teaching method might have an influence (ignoring that students might have been sampled from different populations) on whether a students passes the assignment or not. Most students who have not been taught with *psi* have not passed the assignment, and the students who have been treated with *psi* have passed the assignment in a higher number than students who have been taught with the traditional method.

b)

H_0 : Using *psi* does not affect whether a student passes the assignment or not.

```
psi_glm <- glm(psi_raw$passed ~ psi_raw$psi + psi_raw$gpa, family = binomial)
print(summary(psi_glm)$coefficients)
```

```
##                Estimate Std. Error   z value    Pr(>|z|)
## (Intercept) -11.601565   4.212977 -2.753769 0.005891336
## psi_raw$psi   2.337776   1.040798  2.246138 0.024695156
## psi_raw$gpa   3.063367   1.222868  2.505067 0.012242815
```

The p-value for *psi* is $0.025 < \alpha$ meaning that we can safely reject H_0 concluding that the *psi* usage has an effect on the student's ability to pass the assignment. In addition, the positive sign for the parameter estimate for *psi* ($= 2.338$) also means that when *psi* is used, the linear predictor is increased by the parameter estimate, increasing the odds of passing the assignment by $e^{2.338} = 10.36$.

c)

In order to estimate the probabilities, we assume the model $Pr(Y = 1) = \Psi(\mu + \alpha + \beta X)$ and by extrapolating the coefficients from the summary in *b)* we can use the logistic function $\Psi(x) = \frac{1}{1+e^{-x}}$.

To find the probability that a student passes the assignment if it has **received** *psi* and a gpa of 3 we use:

$x = -11.602 + (2.338 * 1) + (3.063 * 3) \rightarrow x = -0.075 \rightarrow p = \frac{1}{1+e^{0.075}} = 0.481$

To find the probability of a student passing **without** *psi* and a gpa of 3, we use:

$x = -11.602 + (2.338 * 0) + (3.063 * 3) \rightarrow x = -11.602 + (3.063 * 3) \rightarrow x = -2.41 \rightarrow p = \frac{1}{1+e^{2.41}} = 0.082$

d)

As shown in *b)*, using *psi* increases the linear predictor by 2.338 and hence increases the odds of passing the assignment by a factor $e^{2.338} = 10.36$. This means that a student which has been taught using the *psi* method is 10.36 times more likely to pass the assignment than a student who has been taught using the standard method. Given that *gpa* is an independent variable, this number does **not** depend on the gpa of a student.

e)

H_0 : The probability of success of students receiving the *psi* method is the same as that of students who get the traditional method.

```
x <- matrix(c(3, 15, 8, 6), 2, 2)
fisher.test(x)
```

```
##
##  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
```

15 represents the number of students who did not receive *psi* and did **not** show improvement. 6 is the number of students who received *psi* but did **not** show improvement. The *Fisher's exact test* reports a p-value of $0.0265 < \alpha$ meaning that we can safely reject H_0 concluding that there is a significant difference in the probabilities of success for the two different teaching methods and that the __psi__teaching method seems to work.

f)

By studying and analysing the data in *a)*, we have found that the GPA of a student seems to affect also the passing rate of the students. Seeing that the Fisher's exact test for this contingency table does not take it into account, we think that this approach is **wrong**.

g)

- Fisher's exact test is good and accurate (*exact*) for small datasets, especially for 2x2 contingency tables like the one we have.
- It is hard to make predictions with Fisher's exact test, whilst the logistic regression model gives us the coefficients which are useful for making predictions as we did in *c)*.
- Fisher's exact test gives us a more precise p-value to test our null hypotheses compared to that retrieved using the logisitc regression.