

Assignment 3

Group6: Futong Han, Qinghe Gao, Xinyu Fu

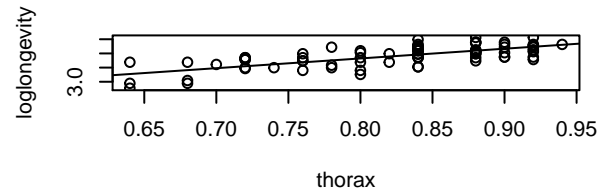
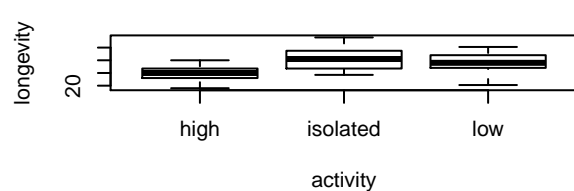
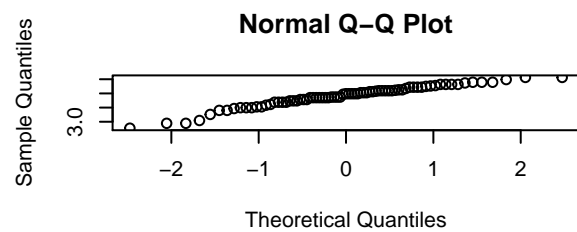
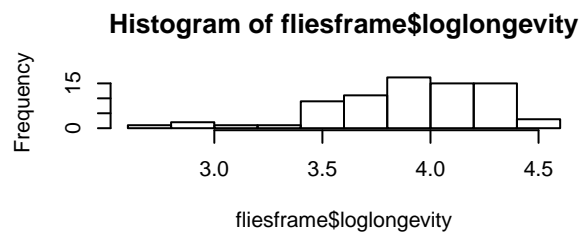
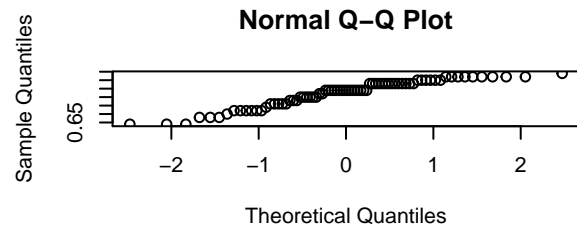
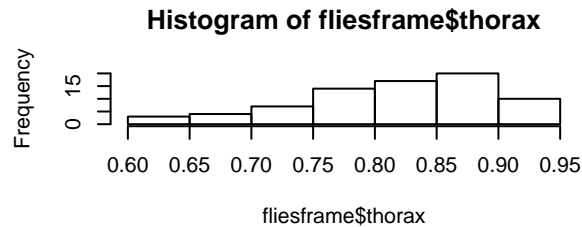
##Exercise 1

We load the data from the data source and add a new column logarithm of longevity in the data frame.

```
flies = read.table("fruitflies.txt", header=TRUE);  
fliesframe = data.frame(thorax=flies$thorax, longevity=flies$longevity, activity=flies$activity, loglongevity=
```

a) We make bloxplot and qqnorm to show the information of the data.

```
par(mfrow=c(3,2));  
hist(fliesframe$thorax);qqnorm(fliesframe$thorax)  
hist(fliesframe$loglongevity);qqnorm(fliesframe$loglongevity)  
plot(longevity~activity, data=fliesframe)  
plot(loglongevity~thorax, data=fliesframe);abline(lm(loglongevity~thorax, data=fliesframe))
```



We can find that the longevity for fruitflies of the activity 'isolated' is the longest and the activity 'high' has the lowest longevity. Also we think that there is a weak linear correlation between loglongevity and thorax. Also, there is a linearity between loglongevity and thorax. Then we perform an anova-test just considering the sexual activity only. And before we use it, we verify the assumption of annova:

```
fliesanova = lm(loglongevity~activity,data=fliesframe)
anova(fliesanova)
```

```
## 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
```

We get a p-value 1.8e-07, therefore H_0 is reject and we can say that sexual activity affects the longevity of the flies.

```
summary(fliesanova)
```

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

According to the estimates longevity increases more when the sexual activity is isolated. High activity = 3.602; Isolated = $3.602 + 0.517 = 4.119$; Low activity = $3.602 + 0.397 = 3.999$ Also we can say that high activity has the shortest longevity and sexual activity decrease longevity. **b)** First we make an assumption that we can use two-way anova test.

Then we will make the 2-way anova test to tell the influence of sexual activity including thorax length.

```
summary(fliesanova2)
```

```
##
## Call:
## lm(formula = loglongevity ~ thorax + activity, data = fliesframe)
##
## 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
```

We obtain a p-value of $1.1e-13$, therefore we can say that activity has a main effect on longevity when we consider the thorax length.

Then we calculate the average thorax length is 0.82 and we calculate three conditions:

With the values $\text{thorax}_9 = 0.312$ (for the average thorax length) and $\text{isolated-activity} = -0.275$, $\text{low-activity} = 0.186$ and $\text{high-activity} = -(-0.275 + 0.186) = 0.089$, then we calculate the estimates for flies with average thorax in three conditions:

$$Y_{\sim\text{isolated}} = (0.82 \cdot 2.98) + 1.22 + 0.41 = 4.07$$

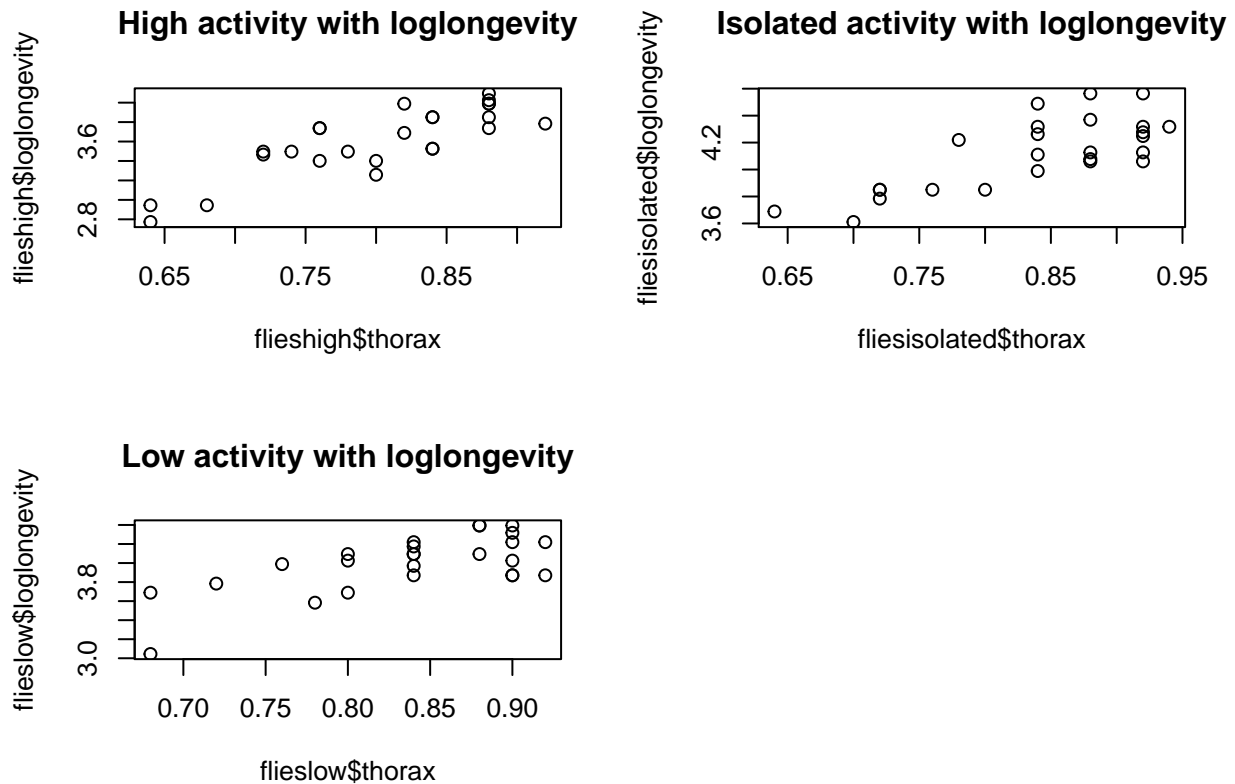
$$Y_{\sim\text{low}} = (0.82 \cdot 2.98) + 1.22 + 0.29 = 3.95$$

$$Y_{\sim\text{high}} = (0.82 \cdot 2.98) + 1.22 = 3.66$$

c)

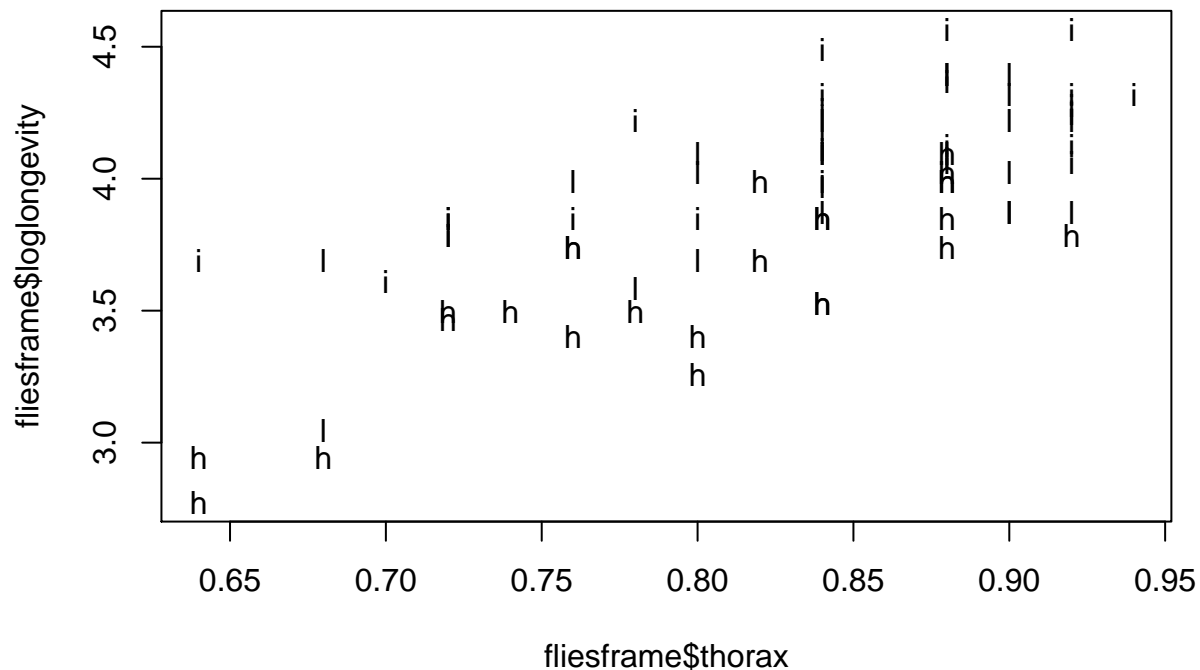
In order to investigate graphically how does thorax length influence the longevity under three conditions, we separate in three different variables the data depending on the sexual activity. From the graph below we could see that longevity increase with the throax.

```
par(mfrow=c(2,2));
plot(flieshigh$thorax,flieshigh$loglongevity,main="High activity with loglongevity")
plot(fliesisolated$thorax,fliesisolated$loglongevity,main="Isolated activity with loglongevity")
plot(flieslow$thorax,flieslow$loglongevity,main="Low activity with loglongevity")
```



Also, we can see that longevity increase with the thorax.

```
plot(fliesframe$loglongevity~fliesframe$thorax,pch=as.character(fliesframe$activity))
```



We consider that thorax will influence the longevity, its dependence on activity is not so clear. Therefore we apply ANCOVA. Using 'drop' to get the p-value. According to result, p-values are less than significance 0.05. It confirms our analysis before that both activity and thorax will influence the longevity.

```
## 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

##
## Call:
## lm(formula = loglongevity ~ thorax + activity, data = fliesframe)
##
## 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
```

From the summary below we could see that p-values for 'isolated:thorax' and 'low:thorax' are bigger than significance level 0.05, therefore we do not reject there is no difference on thorax's dependence under three activities. So the dependence is similar under all three conditions of sexual activity. And also p-value for interaction is 0.15, which means that we accept the H0 there is no interaction between activity and thorax.

```
inter = lm(loglongevity~activity*thorax,data=fliesframe)
anova(inter)
```

```
## Analysis of Variance Table
##
## Response: loglongevity
##           Df Sum Sq Mean Sq F value    Pr(>F)
## activity      2 3.6665  1.8332 45.7687 2.228e-13 ***
## thorax        1 3.8786  3.8786 96.8327 9.020e-15 ***
## activity:thorax 2 0.1542  0.0771  1.9251  0.1536
## Residuals     69 2.7638  0.0401
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(inter)
```

```
##
## Call:
## lm(formula = loglongevity ~ activity * thorax, data = fliesframe)
##
## Residuals:
```

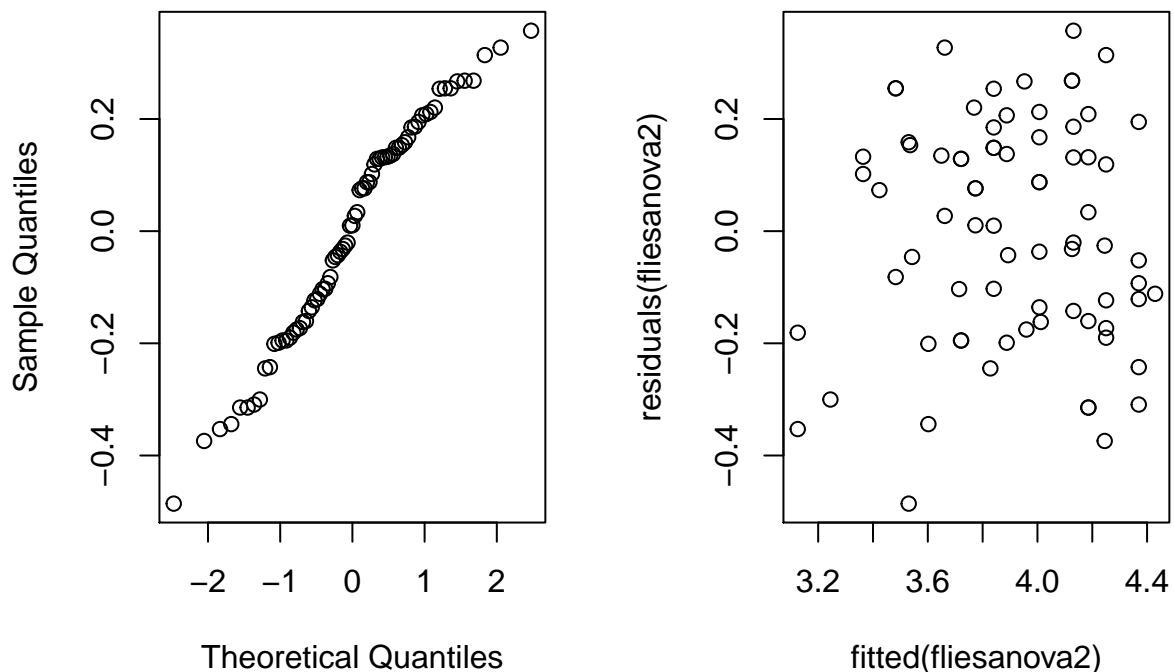
```
##      Min      1Q   Median      3Q      Max
## -0.49803 -0.15920 -0.00031  0.14624  0.35984
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.4372     0.2535   5.669 3.08e-07 ***
## activity1        -0.8394     0.3505  -2.395  0.0194 *
## activity2         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
```

d) We think we need to take thorax into considered because we learn that thorax will influence the longevity of fruitflies. But at the same time we think the first analysis is not wrong. We have to try the individual factor at the beginning and we don't know if thorax affect the longevity.

e) In QQ plot we conclude that it has normality. And from the residuals versus we conclude that there is no sign of heteroscedasticity.

```
par(mfrow=c(1,2));
qqnorm(residuals(fliesanova2))
plot(fitted(fliesanova2),residuals(fliesanova2))
```

Normal Q-Q Plot



f) We do the same anova test using longevity without logarithm as response variable. Even though the residual is normal distribution and variance is homogeneous. But we can see that at the summary part longevity become negative, which means we can not use original data and it is wise to use logarithm of the number of days.

```

fliesanova4 = lm(longevity~thorax+activity, data=fliesframe)
drop1(fliesanova4, test="F")

## Single term deletions
##
## Model:
## longevity ~ thorax + activity
##           Df Sum of Sq   RSS   AIC F value    Pr(>F)
## <none>                 7673 355.10
## thorax    1   7686.8 15360 405.15  71.127 2.624e-12 ***
## activity  2   4966.7 12640 388.53  22.979 2.016e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

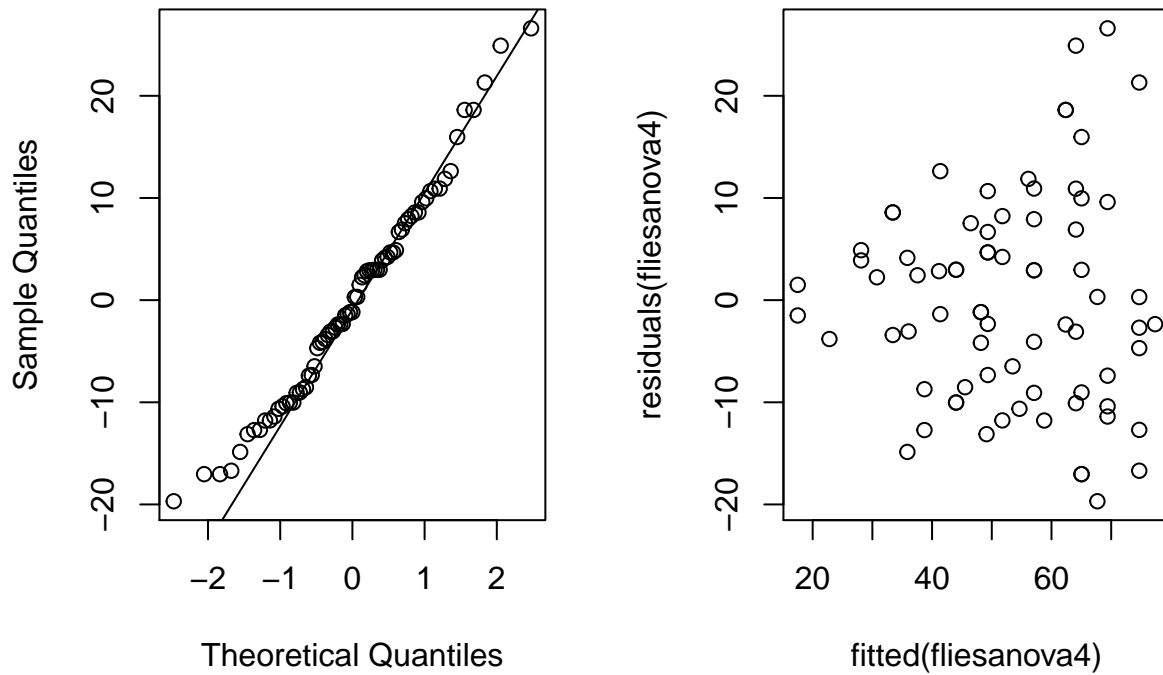
contrasts(fliesframe$activity)=contr.sum
fliesanova4 = lm(longevity~thorax+activity, data=fliesframe)
summary(fliesanova4)

##
## Call:
## lm(formula = longevity ~ thorax + activity, data = fliesframe)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.688  -8.622  -1.176   6.790  26.605
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -56.335     13.021  -4.326 4.87e-05 ***
## thorax        132.618     15.725   8.434 2.62e-12 ***
## activity1     -11.040      1.741  -6.341 1.85e-08 ***
## activity2       9.026      1.707   5.287 1.31e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.4 on 71 degrees of freedom
## Multiple R-squared:  0.6749, Adjusted R-squared:  0.6611
## F-statistic: 49.12 on 3 and 71 DF,  p-value: < 2.2e-16

par(mfrow=c(1,2))
qqnorm(residuals(fliesanova4));qqline(residuals(fliesanova4))
plot(fitted(fliesanova4), residuals(fliesanova4))

```

Normal Q-Q Plot



Exercise 2

##a

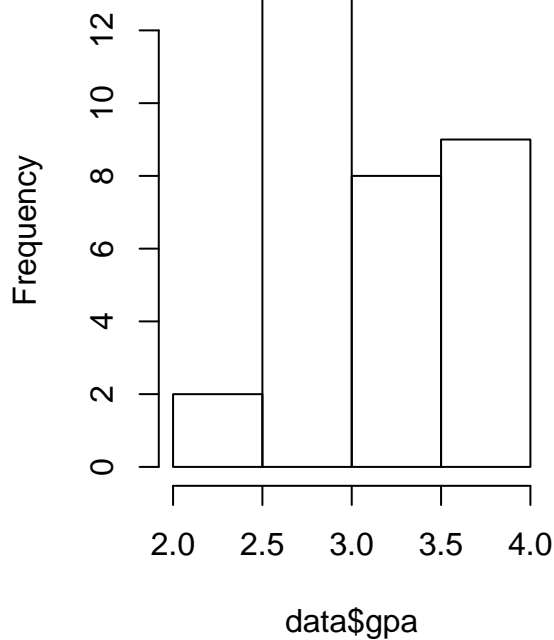
To study the summary of data, we plotted the GPA distribution as a histogram. It turns out more students have a GPA between 2.0 and 3.0 than in other ranges.

We also plotted the quantile-quantile distribution of the GPA data. The qqnorm figure shows the GPA is likely to have a normal distribution.

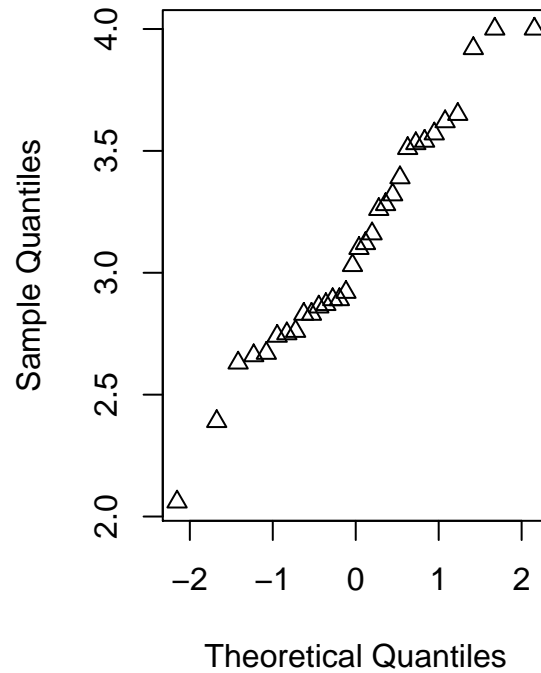
To study the relation between the students who passed the exam and students who received psi instruction, we calculate the contingency table of them.

```
data = read.table("psi.txt",header=TRUE)
par(mfrow=c(1,2))
hist(data$gpa,breaks = 6,pch=1)
qqnorm(data$gpa,pch=2)
```


Histogram of data\$gpa



Normal Q-Q Plot



```
tot=xtabs(~passed+psi,data=data)
tot
```

```
##      psi
## passed 0  1
##      0 15  6
##      1  3  8
```

```
#b
```

To fit the data with a logistic regression model, we assume the binomial form of logistic regression is

$$\log\left(\frac{Pr(Y=1)}{Pr(Y=0)}\right) = \mu + \alpha_i + \beta_i X_i$$

for $i = 1, 2, 3 \dots N$.

We used the generalised linear model (glm) with both explanatory variables to fit the data. And we obtained the fitted logistic model as

$$Pr(Y=1) = \Phi(-11.602 + 2.338X_{psi} + 3.063X_{gpa})$$

for $\Phi = 1/(1 + e^{-x})$

We then perform the test for the $H_0: \alpha_i = 0$ and $\beta_i = 0$ for $i = 1, 2, 3 \dots N$. As a result, we reject the null hypothesis with obtained $p_{psi} = 0.024 < 0.05 = \alpha$ and $p_{gpa} = 0.01224 < 0.05 = \alpha$.

In conclusion, the “psi” works for the performance of students who passed the exam in our test.

```
data = read.table("psi.txt",header=TRUE)
psiglm=glm(passed~psi+gpa,data=data,family=binomial)
summary(psiglm)
```

```
##
```

```
## Call:
## glm(formula = passed ~ psi + gpa, 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 **
## psi           2.338       1.041   2.246  0.02470 *
## gpa           3.063       1.223   2.505  0.01224 *
## ---
## 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
```

To estimate the probability that a student having $gpa = 3$, $psi = 1$ passes the assignment and a student who does not receive psi as $psi = 0$, we put $\beta_{psi} = 0$, $\beta_{psi} = 1$ with $\beta_{gpa} = 3$ respectively to fitted logistic model

$$Pr(Y = 1) = \Phi(-11.602 + 2.338X_{psi} + 3.063X_{gpa})$$

for $\Phi = 1/(1 + e^{-x})$

Thus, we obtained $Pr(passed = 1) = 0.08230$ for a student haveing $gpa = 3$, $psi = 0$ and $Pr(passed = 1) = 0.4815$ for students a student haveing $gpa = 3$, $psi = 1$.

```
predict(psiglm,data.frame(psi=0,gpa=3),"response")
```

```
##           1
## 0.08230274
```

```
predict(psiglm,data.frame(psi=1,gpa=3),"response")
```

```
##           1
## 0.4815864
```

```
#d
```

To estimate the relative change in odds as $\frac{odds_{psi=1}}{odds_{psi=0}}$, we firstly rewrite the fitted model with odds being result

$$odds = \log\left(\frac{Pr(Y = 1)}{Pr(Y = 0)}\right) = -11.602 + 2.338X_{psi} + 3.063X_{gpa}$$

where $x_{psi} = 1$ stands for students having “psi” instructions and $x_{psi} = 0$ stands for students having NO “psi” instructions.

After items arrangements and calculation, we can have the relative change of odds as

$$\frac{odds_{psi=1}}{odds_{psi=0}} = \frac{e^{\beta_0 + \beta_{psi} + \beta_{gpa}X_{gpa}}}{e^{\beta_0 + \beta_{gpa}X_{gpa}}} = e^{\beta_{psi}} = 10.36$$

Thus, we conclude the relative change of odds means that adding psi to any arbitrary student can increase his/her odds to pass the exam by a factor of 10.36 rather than the standard method.

The relative change of odds does not depend on “gpa” since

$$\frac{odds_{psi=1}}{odds_{psi=0}} = e^{\beta_{psi}}$$

which only depends on “psi”

```
exp(2.338)
```

```
## [1] 10.36049
```

```
#d
```

The number 15 means there are 15 out of 18 students receiving NO psi did not show improvement.

The number 6 means there are 6 out of 14 students having psi but does not show improvement.

We performed a Fisher exact test with H0: there is no association between improvement and psi instruction. The result shows we reject the null hypothesis with $p = 0.0265 < 0.05$.

Thus, we conclude that there is a significant association between the improvement of students passing the exams and the psi instructions they received.

```
x=matrix(c(3,15,8,6),2,2);
x
```

```
##      [,1] [,2]
## [1,]    3    8
## [2,]   15    6
```

```
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
```

```
#f
```

However, the experiment conducted is not correct due to the reversed positions of psi groups. Because the odds ratio is defined as the ratio of the odds of A in the presence of B and the odds of A in the absence of B. Thus, a reciprocal odds ratio is obtained.

So we performed a new test with swapped psi group, and we obtained the odds ratio as $OR = 6.22$.

Therefore, we conclude that there is a positive correlation between students having psi instructions and improvements to pass the exams.

```
x=matrix(c(8,6,3,15),2,2);
x
```

```
##      [,1] [,2]
## [1,]    8    3
```

```
## [2,]    6   15
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:
##  1.047057 49.595860
## sample estimates:
## odds ratio
##  6.227408
""

#g
```

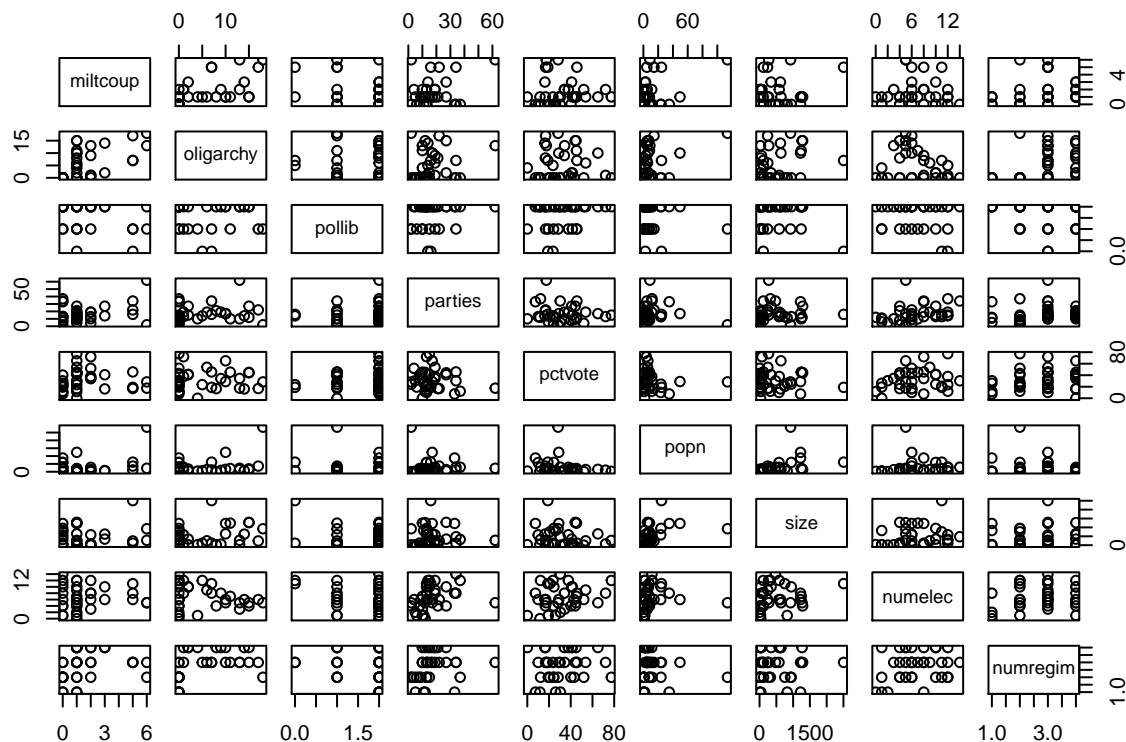
The advantage of the Fisher test is that it has an exact value of P-value, but the logistic regression has an approximation of P-value.

Logistic regression can fit the data having multiple factors with many levels. However, the fisher test usually calculates the p-value on a 2x2 contingency table or with a small sample of data.

Exercise 3

a) First of all, we load the data from the data source and we check if there are any linear correlated factors in the model.

```
africa_data = read.table("africa.txt", header = TRUE)
plot(africa_data)
```



```
round(cor(africa_data),2)
```

```
##           miltcoup oligarchy pollib parties pctvote  popn  size numelec
## miltcoup      1.00      0.61  -0.34   0.31    0.01  0.36  0.14   0.03
## oligarchy     0.61      1.00  -0.13   0.16    0.01  0.35  0.30  -0.19
## pollib       -0.34     -0.13   1.00   0.17    0.20 -0.15 -0.18  -0.17
## parties       0.31      0.16   0.17   1.00   -0.14 -0.08  0.07   0.33
## pctvote       0.01      0.01   0.20  -0.14    1.00 -0.17 -0.10   0.23
## popn          0.36      0.35  -0.15  -0.08   -0.17  1.00  0.43   0.06
## size          0.14      0.30  -0.18   0.07   -0.10  0.43  1.00   0.21
## numelec       0.03     -0.19  -0.17   0.33    0.23  0.06  0.21   1.00
## numregim      0.26      0.46  -0.16   0.19    0.15 -0.14  0.02   0.23
##           numregim
## miltcoup      0.26
## oligarchy     0.46
## pollib       -0.16
## parties       0.19
## pctvote       0.15
## popn         -0.14
## size          0.02
## numelec       0.23
## numregim      1.00
```

We can see that there are no linear correlations from correlation table.

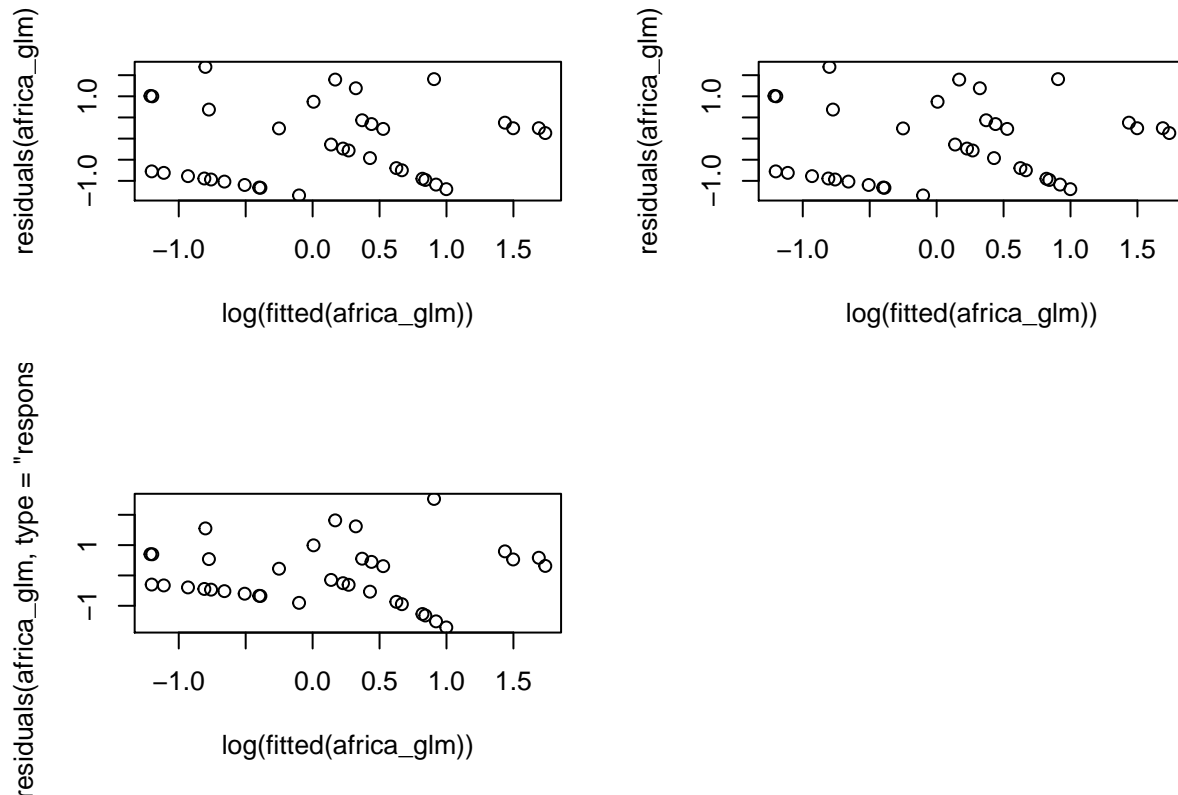
Then with the generalised linear regression model function we use the poisson regression with the following result:

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

```
##
## Call:
## glm(formula = miltcoup ~ oligarchy + pollib + parties + pctvote +
##       popn + size + numelec + numregim, family = poisson, data = africa_data)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.3443  -0.9542  -0.2587   0.3905   1.6953
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.5102693  0.9053301  -0.564  0.57301
## oligarchy    0.0730814  0.0345958   2.112  0.03465 *
## pollib      -0.7129779  0.2725635  -2.616  0.00890 **
## parties      0.0307739  0.0111873   2.751  0.00595 **
## pctvote      0.0138722  0.0097526   1.422  0.15491
## popn         0.0093429  0.0065950   1.417  0.15658
## size        -0.0001900  0.0002485  -0.765  0.44447
## numelec     -0.0160783  0.0654842  -0.246  0.80605
## numregim     0.1917349  0.2292890   0.836  0.40303
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
```

```
##
## Null deviance: 65.945 on 35 degrees of freedom
## Residual deviance: 28.668 on 27 degrees of freedom
## AIC: 111.48
##
## Number of Fisher Scoring iterations: 6
```

```
par(mfrow=c(2,2));
plot(log(fitted(africa_glm)),residuals(africa_glm))
plot(log(fitted(africa_glm)),residuals(africa_glm))
plot(log(fitted(africa_glm)),residuals(africa_glm, type="response"))
```



We find that oligarchy, pollib and parties significantly estimate (or have a linear relation with) the amount of successful military coups. And for the response residuals even though it increases with the (logarithm) of the fitted values we see the specific structure. **b)** In the step down method we have removed the following factors in order: numelec > numregim > size > popn > pctvote.

Then the miltcoup is: $0.251 + 0.0926 * \text{oligarchy} - 0.574 * \text{pollib} + 0.0220 * \text{parties} + \text{error}$. With this process we went from a R squared value of 0.565 to 0.502, but reduced the formula from eight factors to three.

```
options(digits = 3)
africaglm2=glm(miltcoup~oligarchy+pollib+parties, family=poisson,data=africa_data)
with(summary(africaglm2), 1 - deviance/null.deviance)
```

```
## [1] 0.502
```

```
summary(africaglm2)
```

```
##
## Call:
## glm(formula = miltcoup ~ oligarchy + pollib + parties, family = poisson,
## data = africa_data)
```

```
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.358  -1.042  -0.286   0.628   1.752
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.25138    0.37269   0.67   0.500
## oligarchy    0.09262    0.02178   4.25 2.1e-05 ***
## pollib      -0.57410    0.20438  -2.81  0.005 **
## parties      0.02206    0.00896   2.46  0.014 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##      Null deviance: 65.945  on 35  degrees of freedom
## Residual deviance: 32.856  on 32  degrees of freedom
## AIC: 105.7
##
## Number of Fisher Scoring iterations: 5
```

And then we verify the assumption of Poisson regression. The response residuals clearly increase with the (logarithm) of the fitted values, as expected under a Poisson model.

```
par(mfrow=c(2,2));
plot(log(fitted(africaglm2)),residuals(africaglm2))
plot(log(fitted(africaglm2)),residuals(africaglm2))
plot(log(fitted(africaglm2)),residuals(africaglm2, type="response"))
```

