assignment3

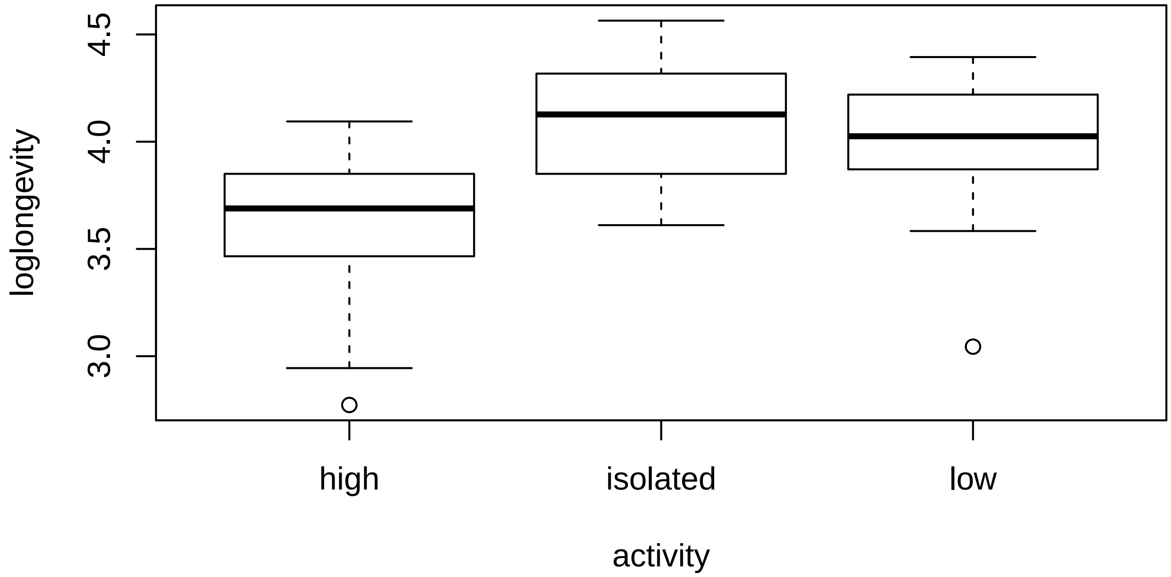
nihat uzunalioglu - 2660298, emiel kempen - 2640580, saurabh jain - 2666959

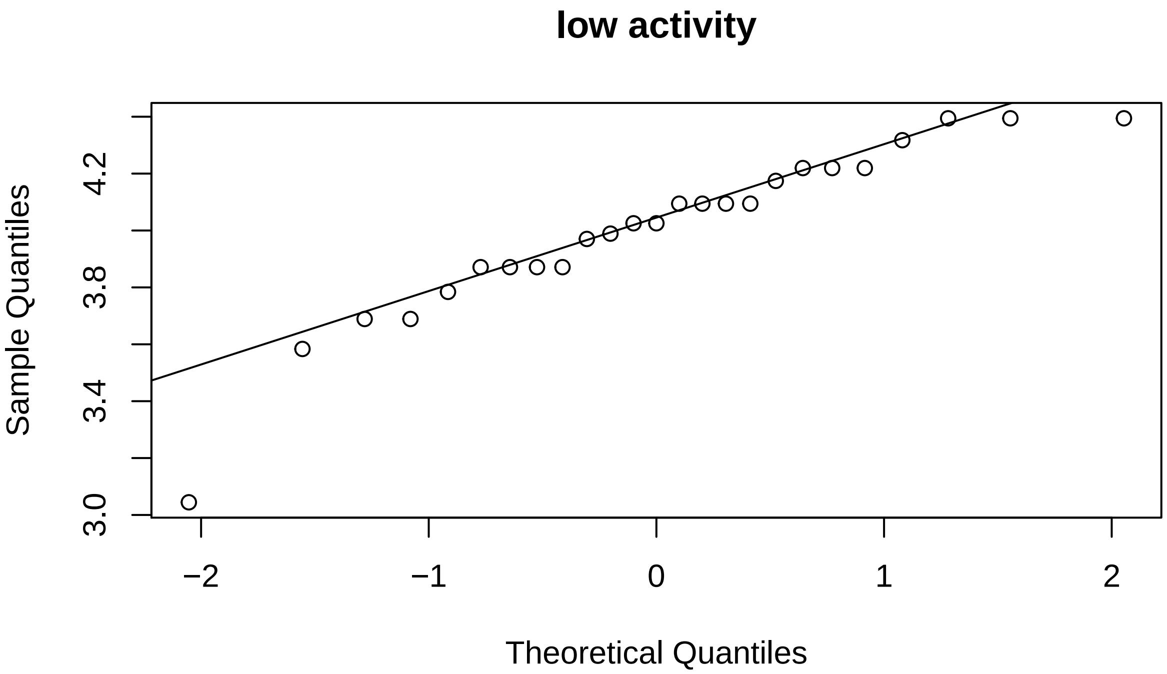
3/15/2020

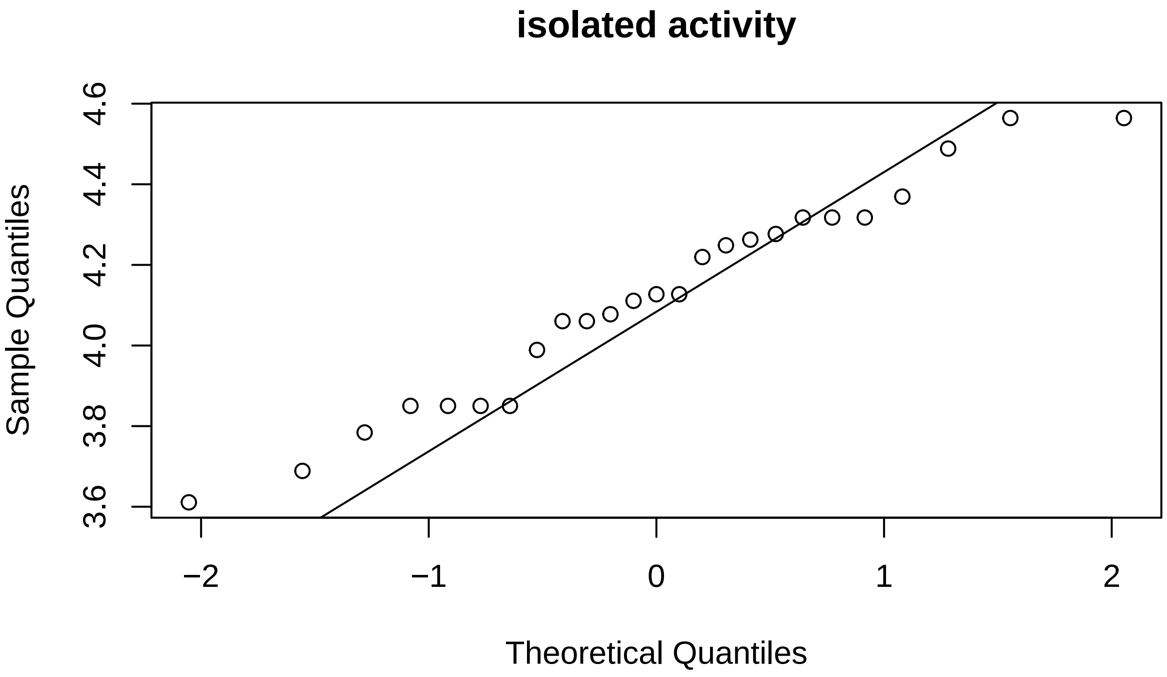
Exercise 1

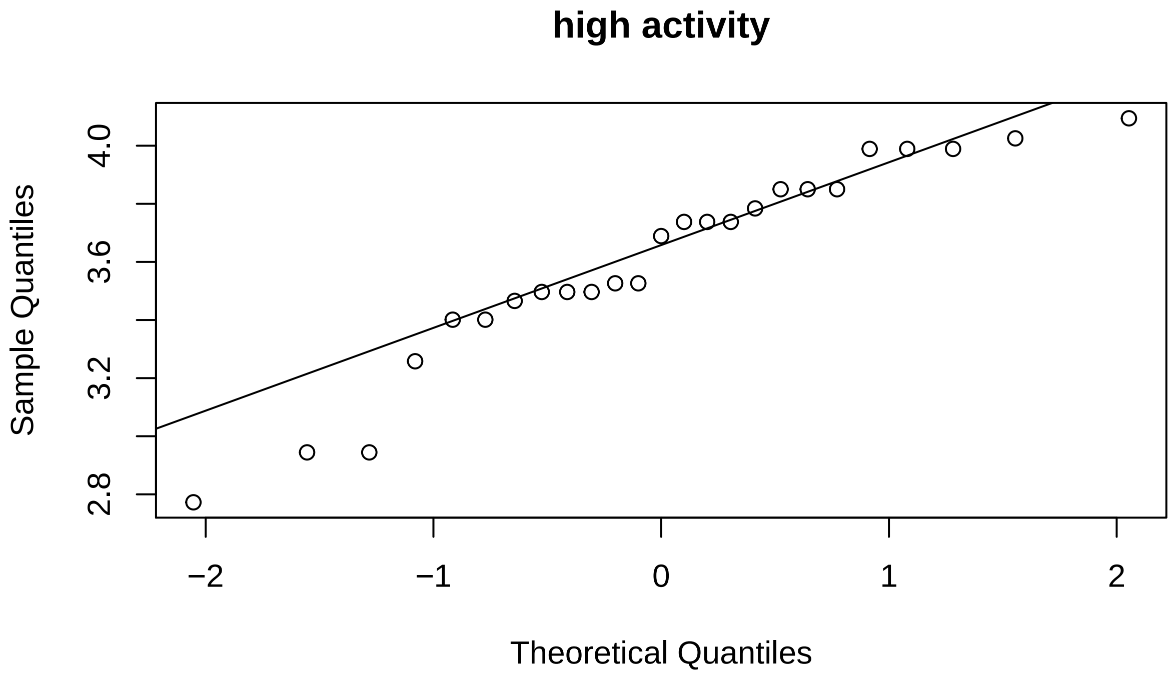
```
# Add loglongevity to the data-frame
# Use it as a response variable (Y)
fruitflies$loglongevity = log(fruitflies$longevity)
```

Section A




low activity
Sample Quantiles
3.0
3.4
3.8
4.2
−2
−1
0
1
2
Theoretical Quantiles


isolated activity
Sample Quantiles
3.6
3.8
4.0
4.2
4.4
4.6
−2
−1
0
1
2
Theoretical Quantiles

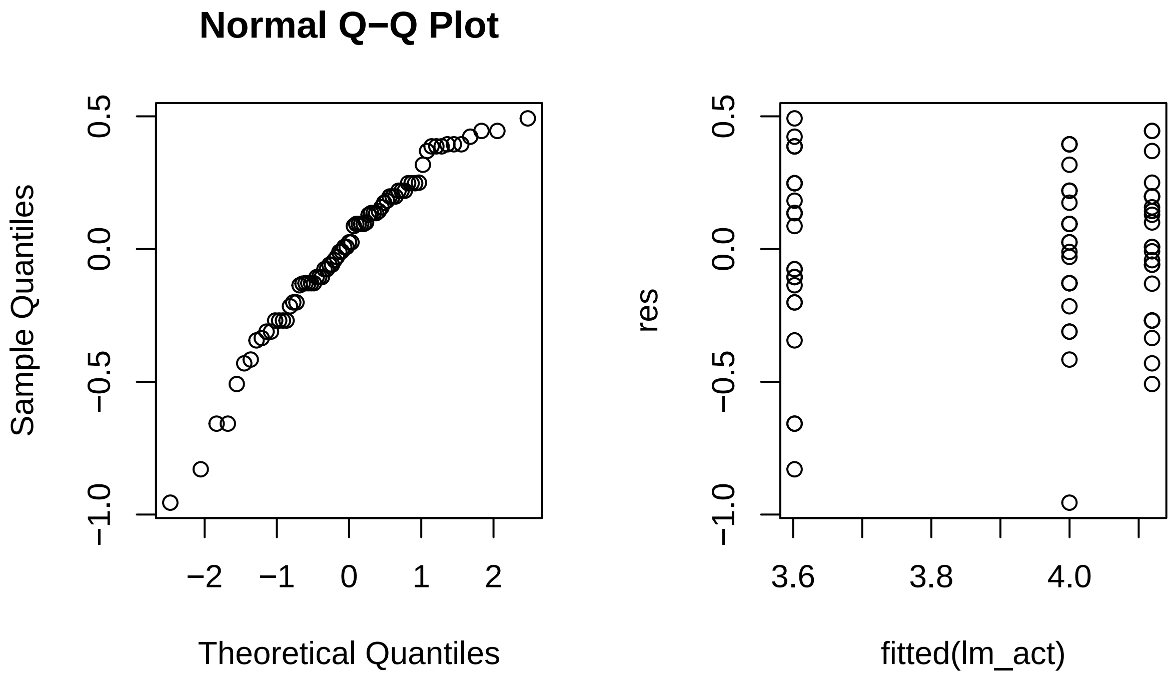


- From the boxplot, we can observe medians are not the same among activities and there is also an outlier both in high and low activity. From QQ-plots, we observed normality on low activity with 2 outliers, whereas isolated and high activities don't have normality.

```
## Analysis of Variance Table
##
## Response: loglongevity
##           Df Sum Sq Mean Sq F value  Pr(>F)
## activity   2   3.67   1.833    19.4 1.8e-07
## Residuals 72   6.80   0.094
```

```
## $coefficients
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.602     0.0614   58.62 1.65e-62
## activityisolated    0.517     0.0869    5.95 8.82e-08
## activitylow         0.398     0.0869    4.58 1.93e-05
```

- We reject the null hypothesis of ANOVA test which means there is significant difference among groups and also we can observe the same outcome from the summary coefficient table. According to the estimations of longevity, it increases more when the sexual activity is isolated, then it follows as low and high. To find out the αs;
  - $\alpha_{high} = \mu_1 = 3.602$,
  - $\alpha_{isolated} = \mu_2 - \mu_1 = 0.517 ==> \mu_2 = 4.119$,
  - $\alpha_{low} = \mu_3 - \mu_1 = 0.398 ==> \mu_3 = 4$



- QQ plot does not provide good results with regards to the normality. Shapiro - Wilk test also supplies the same outcome with 0.009 value since we rejected null hypothesis H_0 (because it is lower than 0.05) which means residuals are not normally distributed and that is a sign which tells something wrong about our model, even though, we created our model with log of longevity (a transformation to longevity for the sake of the model beforehand).

Section B

```
## [1] 0.825
```

```
## Analysis of Variance Table
##
## Response: loglongevity
##           Df Sum Sq Mean Sq F value Pr(>F)
## thorax    13   5.99   0.461    12.9  8e-13
## activity   2   2.37   1.187    33.4  2e-10
## Residuals 59   2.10   0.036
```

```
## $coefficients
##                   Estimate Std. Error  t value Pr(>|t|)
## (Intercept)        2.98120     0.1107 26.93001 7.68e-35
## thorax0.68         0.00141     0.1583  0.00889 9.93e-01
## thorax0.7          0.16742     0.2214  0.75616 4.53e-01
## thorax0.72         0.43197     0.1345  3.21118 2.14e-03
## thorax0.74         0.51530     0.2187  2.35592 2.18e-02
## thorax0.76         0.59647     0.1382  4.31561 6.16e-05
## thorax0.78         0.50954     0.1554  3.27876 1.75e-03
## thorax0.8         0.47891     0.1362  3.51541 8.51e-04
## thorax0.82        0.85773     0.1733  4.94811 6.58e-06
## thorax0.84        0.74175     0.1219  6.08705 9.27e-08
## thorax0.88        0.90173     0.1208  7.46272 4.45e-10
## thorax0.9         0.77032     0.1444  5.33390 1.60e-06
## thorax0.92        0.79276     0.1291  6.14228 7.50e-08
## thorax0.94        0.87399     0.2214  3.94748 2.13e-04
## activityisolated  0.46230     0.0594  7.77695 1.31e-10
## activitylow       0.36501     0.0620  5.89107 1.96e-07
```

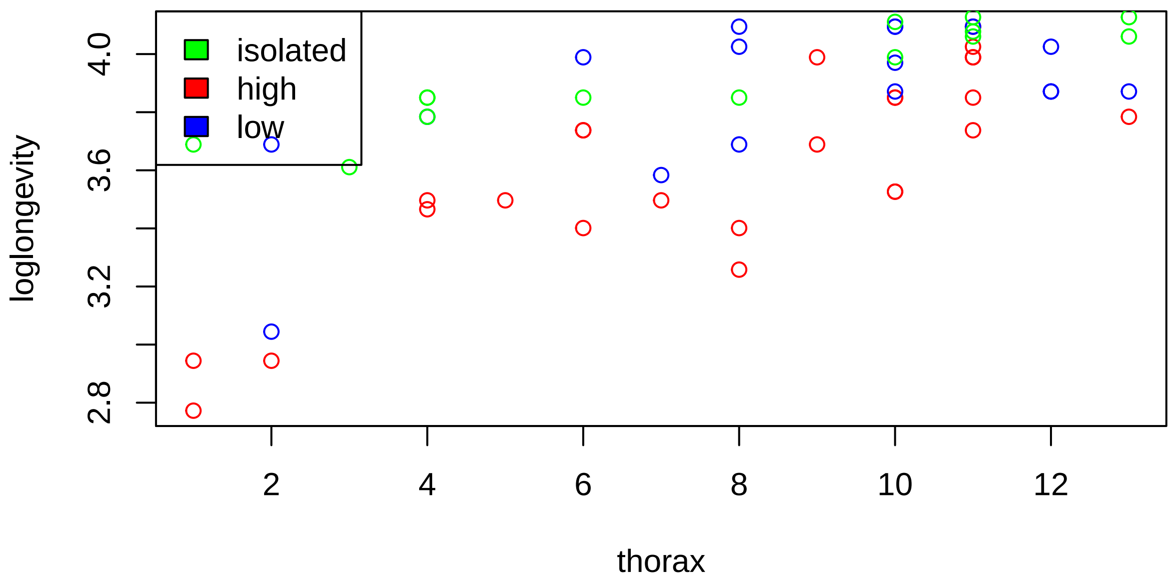
- According to the two-way ANOVA results, we observe that thorax (α_i) and activity (β_j) have both main effects on longevity in the additive model since we rejected the null hypothesis.
- From the summary coefficients, we observe that all activity factors increase longevity since all of them are positive.
- We obtain thorax average as 0.825

```
## Analysis of Variance Table
##
## Response: loglongevity
##           Df Sum Sq Mean Sq F value Pr(>F)
## thorax    13   5.99   0.461    12.9  8e-13
## activity   2   2.37   1.187    33.4  2e-10
## Residuals 59   2.10   0.036
```

```
## $coefficients
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.8026     0.0317 119.973 3.45e-72
## thorax1      -0.5457     0.1070  -5.101 3.76e-06
## thorax2      -0.5443     0.1082  -5.028 4.92e-06
## thorax3      -0.3782     0.1807  -2.093 4.06e-02
## thorax4      -0.1137     0.0783  -1.451 1.52e-01
## thorax5      -0.0304     0.1803  -0.168 8.67e-01
## thorax6       0.0508     0.0850   0.598 5.52e-01
## thorax7      -0.0361     0.1057  -0.342 7.34e-01
## thorax8      -0.0668     0.0791  -0.844 4.02e-01
## thorax9       0.3121     0.1314   2.375 2.08e-02
## thorax10      0.1961     0.0556   3.529 8.15e-04
## thorax11      0.3561     0.0564   6.318 3.82e-08
## thorax12      0.2247     0.0876   2.563 1.29e-02
## thorax13      0.2471     0.0685   3.606 6.40e-04
## activity1    -0.2758     0.0348  -7.933 7.11e-11
## activity2     0.1865     0.0349   5.343 1.54e-06
```

- As contr.sum equals to zero, we calculated activity3 (high) as -(-0.2758 + 0.1865) = 0.089. So, to have the estimates for flies with average thorax, we move on with the values thorax9 = 0.3121 (for the average thorax length), and all activity factors.
  - $Y_{isolated,thorax9}$ = 3.8026 + 0.3121 - 0.2758 = 3.839,
  - $Y_{low,thorax9}$ = 3.8026 + 0.3121 + 0.1865 = 4.301,
  - $Y_{high,thorax9}$ = 3.8026 + 0.3121 + 0.089 = 4.204

Section C



- There is an increment in longevity as thorax length increases. Moreover, flies which were in isolated sexual activity seem to have the longer than the others. Additionally, it follows as low and isolated activity factors, respectively. But to be sure, we also obtain results from ANOVA additive model.

```
## Analysis of Variance Table
##
## Response: loglongevity
##           Df Sum Sq Mean Sq F value  Pr(>F)
## thorax     1   5.41    5.41     132 < 2e-16
## activity   2   2.14    1.07      26 3.3e-09
## Residuals 71   2.91    0.04
```

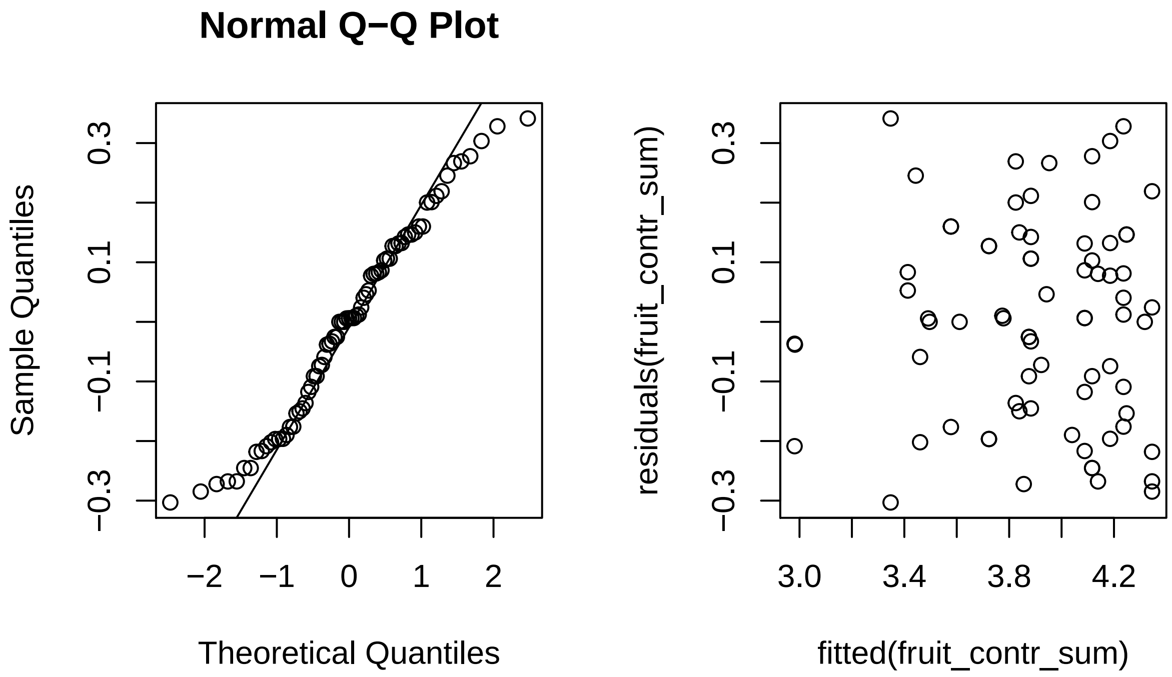
```
## $coefficients
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.0762    0.06758   45.52 2.82e-54
## thorax             0.0674    0.00693    9.72 1.10e-14
## activityisolated   0.4120    0.05832    7.07 8.92e-10
## activitylow        0.2871    0.05843    4.91 5.52e-06
```

- According to the ANOVA test results, we reject null hypothesis for both activity and thorax which leads to the fact that they have significant effect upon longevity. Whereas we can't say we reject null hypothesis for the interaction between activity and thorax.
  - α_1 (high) $= \mu_1 = 3.0762$,
  - α_2 (isolated) $= \mu_2 - \mu_1 = 3.0762 + 0.4120 = 3.488$,
  - α_3 (low) $= \mu_2 - \mu_1 = 3.0762 + 0.2871 = 3.363$,
  - When we look the the results the highest effect is supplied by low activity, then isolated and lastly high sexual condition.

Section D

- We would prefer without thorax parameter since there is no real interaction between activity and thorax. Moreover, as in the beginning of the question, it says that experimenters randomly chose the sexual activity upon flies will going to experience and most importantly, thorax is considered to be added later on which does not seem a reliable factor for this testing.

Section E



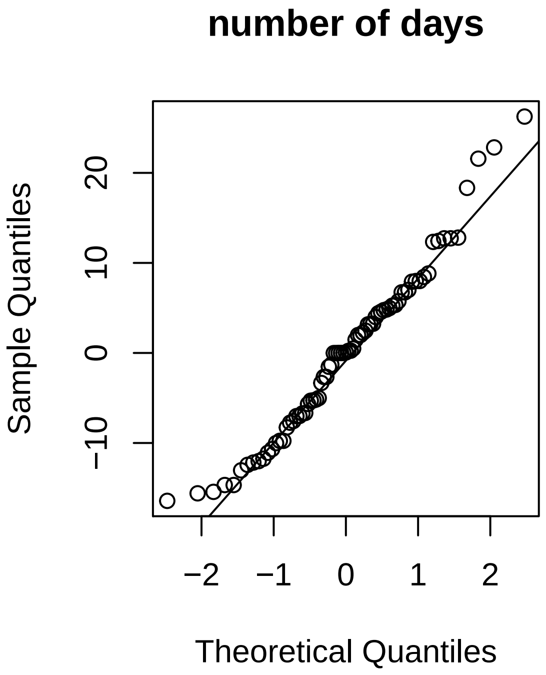
- The normality seems doubtful in QQ-plot and we did not observe heteroscedasticity in the scatter plot as they mostly are accumulated on the right side of the plot.

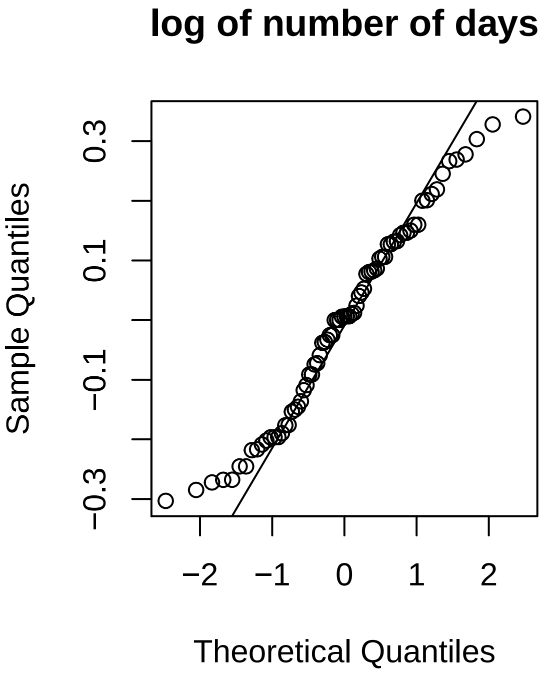
```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##
##  studentized Breusch-Pagan test
##
## data:  fruit_contr_sum
## BP = 22, df = 15, p-value = 0.1
```

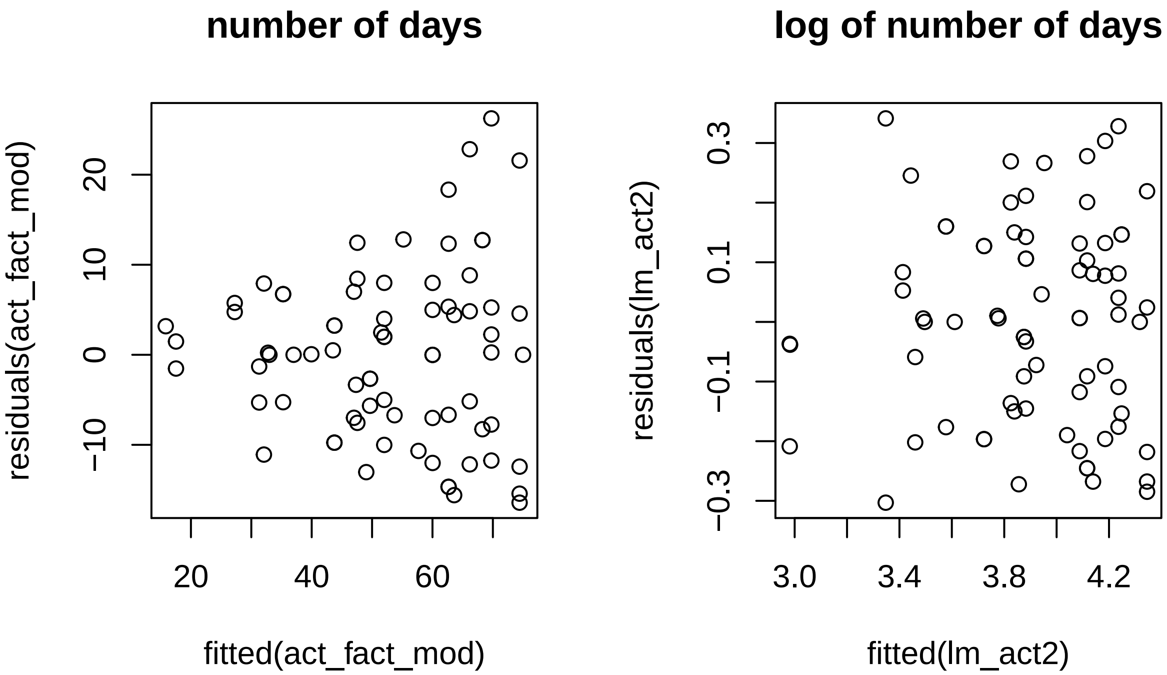
- We also don’t reject the null hypothesis of Breusch-Pagan test which is the error variances are all equal and that is a proof of non-heteroscedasticity among the model.

Section F

```
## Single term deletions
##
## Model:
## fruit$longevity ~ fruit$thorax + fruit$activity
##                 Df Sum of Sq   RSS AIC
## <none>                        6561 367
## fruit$thorax   13      8799 15360 405
## fruit$activity  2      5377 11938 408
```







- Normality is doubtful for both of the QQ-plots. But residuals of the number of days seem to be more normal distributed than the log of them.

```
##
##  studentized Breusch-Pagan test
##
## data:  act_fact_mod
## BP = 19, df = 15, p-value = 0.2
```

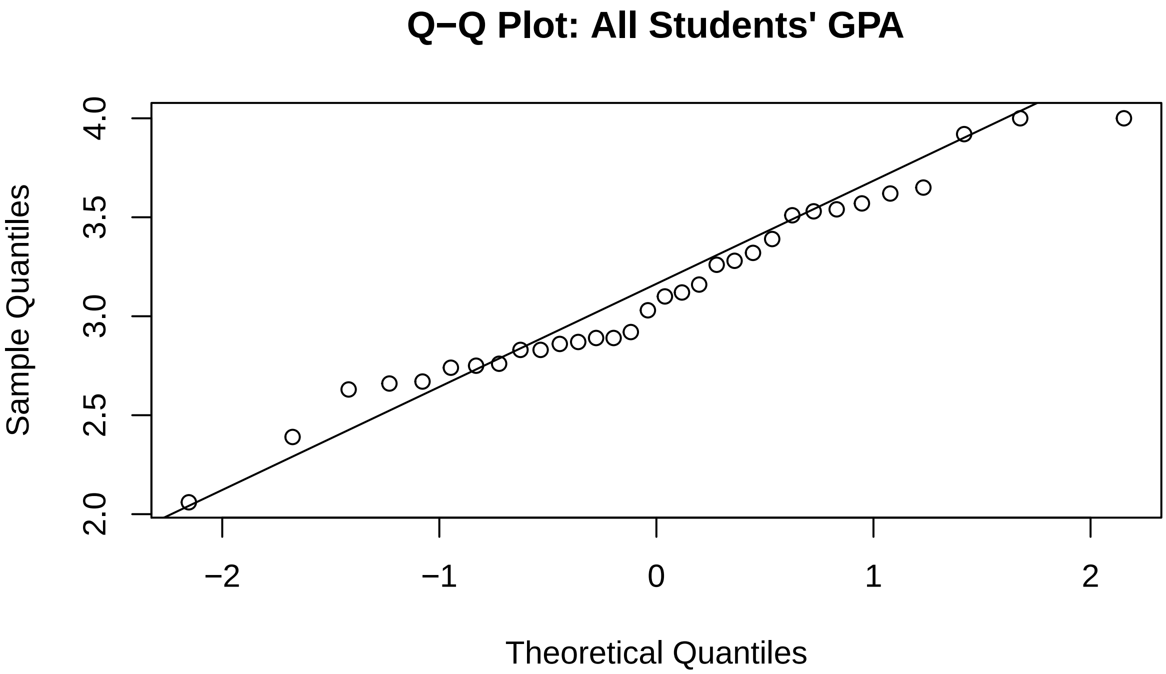
```
##
##  studentized Breusch-Pagan test
##
## data:  lm_act2
## BP = 22, df = 15, p-value = 0.1
```

- According to the Breush-Pagan tests for both numerical (first test) and log of longevity values (second test), we don't reject, and therefore, confirm non-heteroscedasticities for both of them. Since we have better QQ-plot and greater p-value from Breush-Pagan test, we conclude as the real numerical values for longevity is better than logarithmic for the model.

Exercise 2

a)

First, we load the data and check the normality of the GPA data.

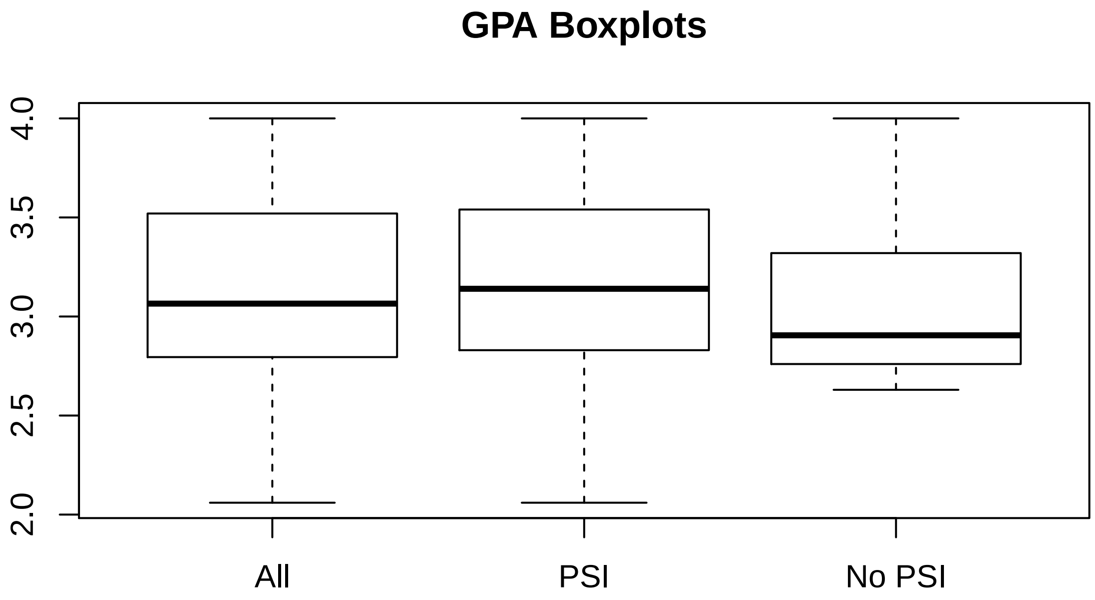


From the QQ-plot the normality can not be assumed from the plot, so we perform a Shapiro Test to make sure. The test outputs a p-value of 0.492147662355498, which means the GPA data is normally distributed.

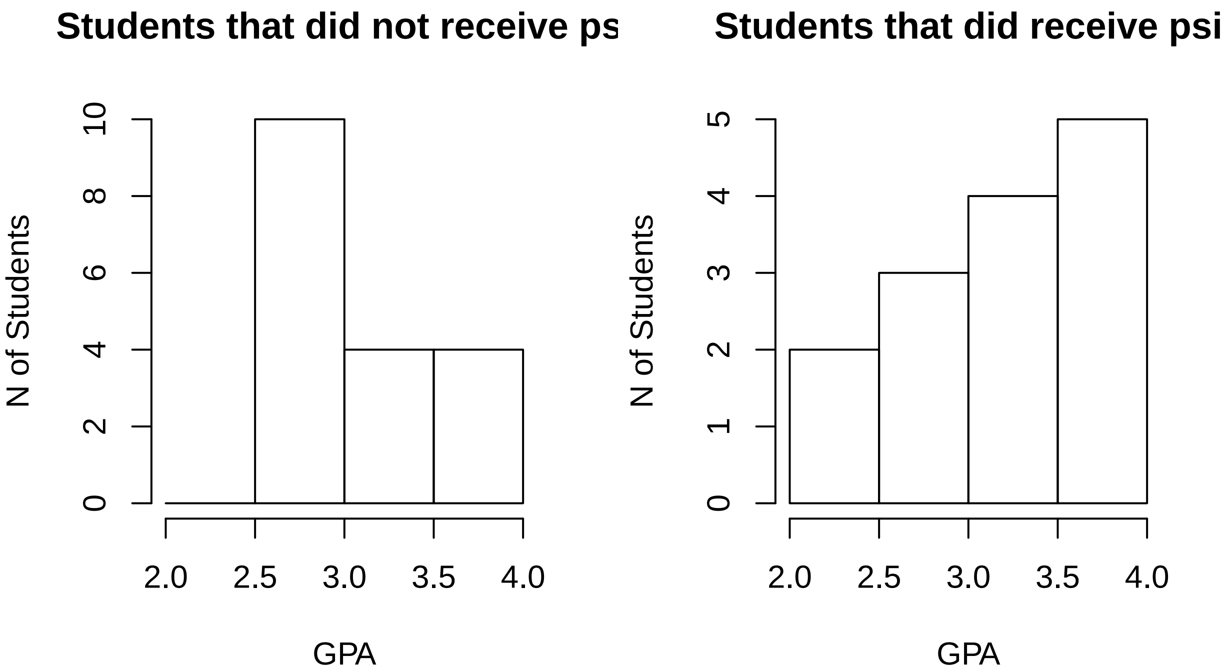
Then, we divide the dataset into two groups based on whether they received psi, to be able to study the data and give summaries.

```
no_psi = psi[which(psi$psi == "0"),]
yes_psi = psi[which(psi$psi == "1"),]
```

The splitted data can then be studied from the following boxplot.



We can see that the mean GPA of students with no psi is lower compared to the mean GPA of students with psi, but the latter has a higher variance.



From the boxplots it clearly shows that students that did receive psi are more lenient towards the higher end of the GPAs, for students without psi the data is more concentrated with GPAs around the 2.5 - 3.0 mark.

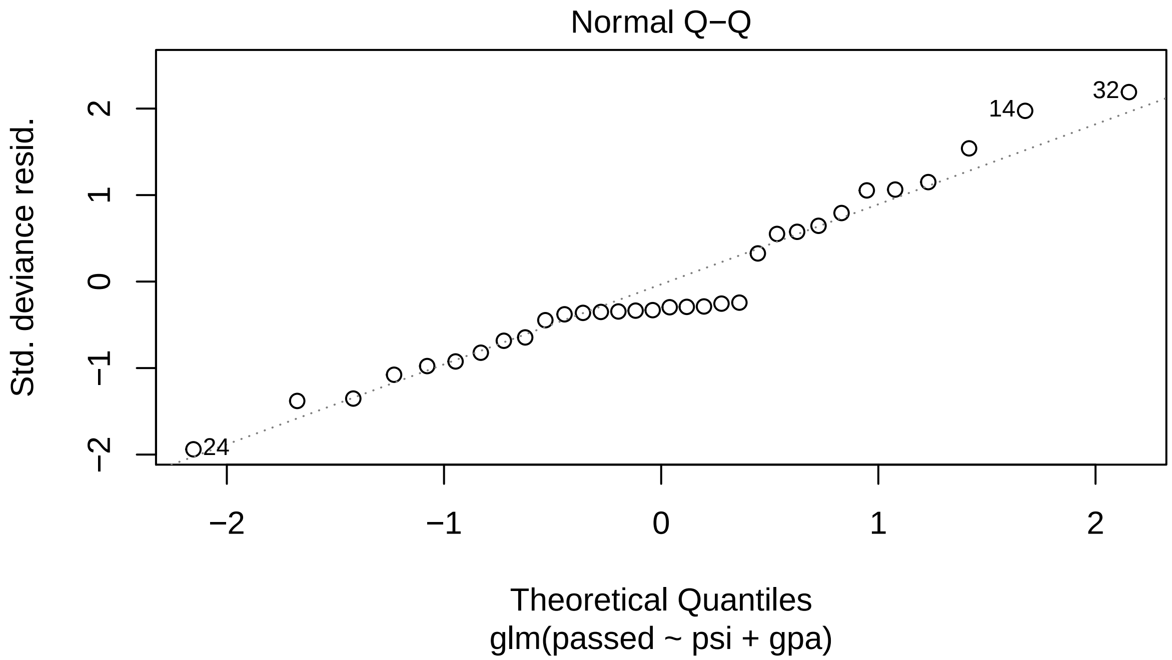
b)

Fitting a logistic regression model.

```
psiglm=glm(passed~psi+gpa,data=psi,family=binomial)
summary(psiglm)$coefficients
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -11.60       4.21   -2.75  0.00589
## psi             2.34       1.04    2.25  0.02470
## gpa             3.06       1.22    2.51  0.01224
```

```
plot(psiglm, which=2:2)
```



From the summary follows that the p-value for receiving psi equals 0.012 so we reject the H_0, concluding that receiving psi does work. Also, the parameter estimate of psi of 2.34 shows that that if psi is received, the probability of passing is higher. From the graph it shows that the residuals seem quite normal, which shows that our model works well. It also show that points 24, 14 and 32 are outliers in the data set.

c)

From the summary of the previous question follow the following probabilities:

A student with a GPA equal to 3 who receives psi

- $-11.602 + 2.338 + (3.063 * 3.0) = -0.075$
- Probability: $\frac{1}{1+e^{-(-0.075)}} = 0.481$

A student with a GPA equal to 3 who does not receives psi

- $-11.602 + (3.063 * 3.0) = -2.408$
- Probability: $\frac{1}{1+e^{-(-2.408)}} = 0.083$

The probability of a student that receives psi passing the assigment is higher than the student with a same GPA who did not receive psi.

d)

From the same summary follow the following probabilities.

A student that received psi:

- $-11.602 + 2.34 = -9.26$
- Probability: $\frac{1}{1+e^{-(-9.26)}} = 9.515 \times 10^{-5}$

A student that did not receive psi:

- Probability: $\frac{1}{1+e^{-(-11.602)}} = 9.148 \times 10^{-6}$

The relative change in odds is as follows:

- $9.515 \times 10^{-5} - 9.148 \times 10^{-6} = 8.6 \times 10^{-5}$

This number means that a student's odds increase by 8.6×10^{-5} to pass the assigment when receiving psi. This relative change is not dependent on GPA, because if we leave GPA out of the GLM function, the parameter estimate for psi - the number that is used in this calculation - does not change.

e)

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

In the matrix x, the numbers 15 and 6 equal the number of students that did not show improvements while not receiving and receiving psi respectively. The conclusion is that the H_0 - that states that the probability of passing is the same for the two groups - is rejected due to the p-value of 0.0265032806177556.

f)

Fisher's exact test is meant to compute an exact p-value for 2x2 tables. The test is mainly used for small sample sizes, which is also the case in our example. This, combined with the way the experiment was conducted, makes us believe that this approach is not wrong.

g)

The last approach could make use of Fisher's exact test, which produces an exact p-value, rather than an approximation that only becomes exact as the sample size grows to infinity, as with many other statistical tests, the one used in part b) of this question included.

However, Fisher's test, in contrast to logistic regression, can only be used for relatively small sample sizes (i.e., small counts in the 2x2 table). Therefore it was not possible to conduct this test on the data acquired by the experiment with the first approach of this question.

Exercise 3

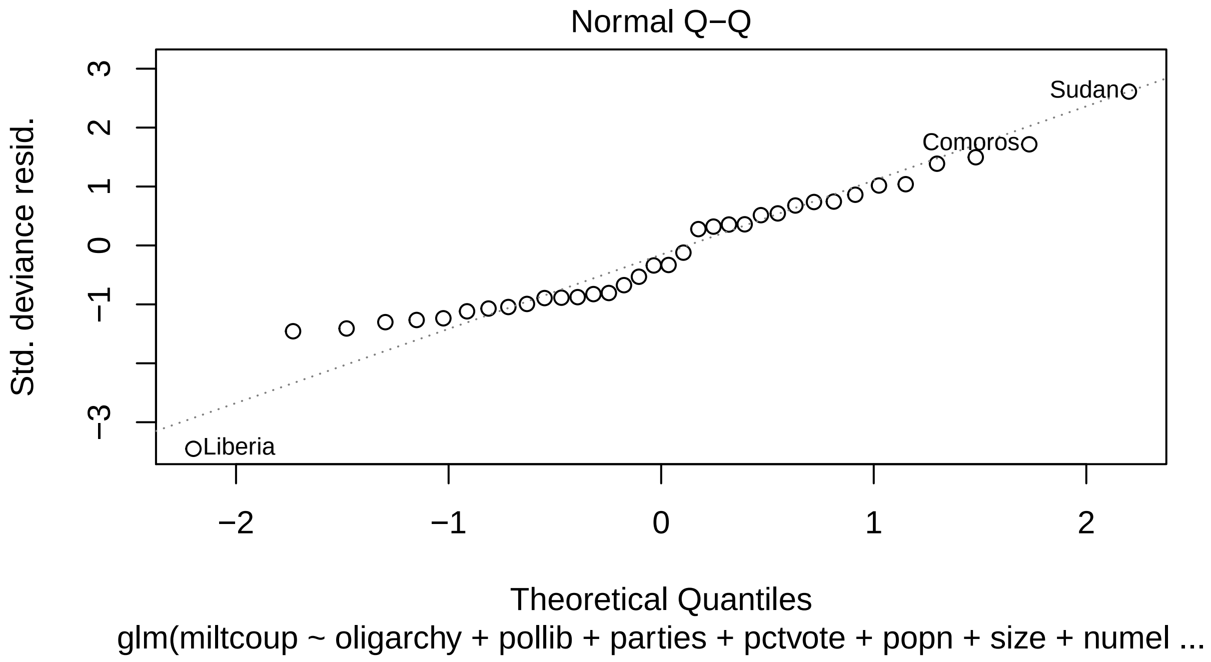
```
africa = read.table(file = 'data/africa.txt', header = TRUE)
```

a) We need to treat explanatory variable `pollib` as a factor variable as it represents three different categories and it is not a numerical variable.

```
africa$pollib = as.factor(africa$pollib)
glm_milt=glm(miltcoup~oligarchy+pollib+parties+pctvote+popn+size+numelec+numregim,family=poisson,data=a
summary(glm_milt)
```

```
##
## Call:
## glm(formula = miltcoup ~ oligarchy + pollib + parties + pctvote +
##     popn + size + numelec + numregim, family = poisson, data = africa)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.508  -0.953  -0.310   0.486   1.646
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.233427   0.997611   -0.23   0.8150
## oligarchy    0.072566   0.035346    2.05   0.0401
## pollib1     -1.103244   0.655811   -1.68   0.0925
## pollib2     -1.690306   0.676650   -2.50   0.0125
## parties      0.031221   0.011166    2.80   0.0052
## pctvote      0.015441   0.010103    1.53   0.1264
## popn         0.010959   0.007149    1.53   0.1253
## size        -0.000265   0.000269   -0.99   0.3244
## numelec     -0.029619   0.069625   -0.43   0.6705
## numregim     0.210943   0.233933    0.90   0.3672
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 65.945  on 35  degrees of freedom
## Residual deviance: 28.249  on 26  degrees of freedom
## AIC: 113.1
##
## Number of Fisher Scoring iterations: 5
```

```
plot(glm_milt, which=2)
```



```
shapiro.test(residuals(glm_milt))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(glm_milt)
## W = 0.9, p-value = 0.04
```

`oligarchy`, `parties`, and `pollib` variables are significant for the model as the p-values are less than 5%. As per the plot, `Liberia`, `Comoros`, and `Sudan` are the outliers in the data. Also, we reject the H_0 hypothesis of 'Shapiro test' and conclude the residuals of the model are not normally distributed, meaning that the model does not work well on the given dataset.

b) In 'Step down' method, we start with all the variables and then reduce the number of variables in our model based on the p-value.

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

```
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.233427   0.997611  -0.234  0.81500
## oligarchy    0.072566   0.035346   2.053  0.04007
## pollib1     -1.103244   0.655811  -1.682  0.09252
## pollib2     -1.690306   0.676650  -2.498  0.01249
## parties      0.031221   0.011166   2.796  0.00517
## pctvote      0.015441   0.010103   1.528  0.12641
## popn         0.010959   0.007149   1.533  0.12531
## size        -0.000265   0.000269  -0.985  0.32444
## numelec     -0.029619   0.069625  -0.425  0.67054
## numregim     0.210943   0.233933   0.902  0.36720
```

As the variable `numelec` has the highest p-value and it is > 0.05, we discard it for the next iteration.

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

```
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.457746   0.860234  -0.532  0.59464
## oligarchy    0.081202   0.028815   2.818  0.00483
## pollib1     -0.964298   0.562094  -1.716  0.08625
## pollib2     -1.514951   0.526944  -2.875  0.00404
## parties      0.029341   0.010310   2.846  0.00443
## pctvote      0.013912   0.009465   1.470  0.14164
## popn         0.009959   0.006725   1.481  0.13862
## size        -0.000269   0.000269  -1.000  0.31710
## numregim     0.180442   0.224117   0.805  0.42075
```

In this iteration, variable `numregim` have the p-value > 0.05, thus we discard it for the next iteration.

```
summary(glm(miltcoup~oligarchy+pollib+parties+pctvote+popn+size,family=poisson,data=africa))[[12]]
```

```
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.041976   0.577410  0.0727 0.942048
## oligarchy    0.089495   0.027044  3.3092 0.000936
## pollib1     -0.967325   0.560560 -1.7256 0.084412
## pollib2     -1.532113   0.523278 -2.9279 0.003412
## parties      0.028817   0.010217  2.8204 0.004796
## pctvote      0.014922   0.009376  1.5914 0.111513
## popn         0.007165   0.005684  1.2604 0.207510
## size        -0.000258   0.000266 -0.9688 0.332621
```

In this iteration, variable `size` have the p-value > 0.05, thus we discard it for the next iteration.

```
summary(glm(miltcoup~oligarchy+pollib+parties+pctvote+popn,family=poisson,data=africa))[[12]]
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.23143    0.52889  -0.438  0.66168
## oligarchy    0.08347    0.02583   3.232  0.00123
## pollib1     -0.68359    0.49582  -1.379  0.16799
## pollib2     -1.32057    0.49027  -2.694  0.00707
## parties      0.02977    0.01031   2.887  0.00388
## pctvote      0.01392    0.00937   1.486  0.13728
## popn         0.00566    0.00548   1.032  0.30204
```

In this iteration, variable `popn` have the p-value > 0.05, thus we discard it for the next iteration.

```
summary(glm(miltcoup~oligarchy+pollib+parties+pctvote,family=poisson,data=africa))[[12]]
```

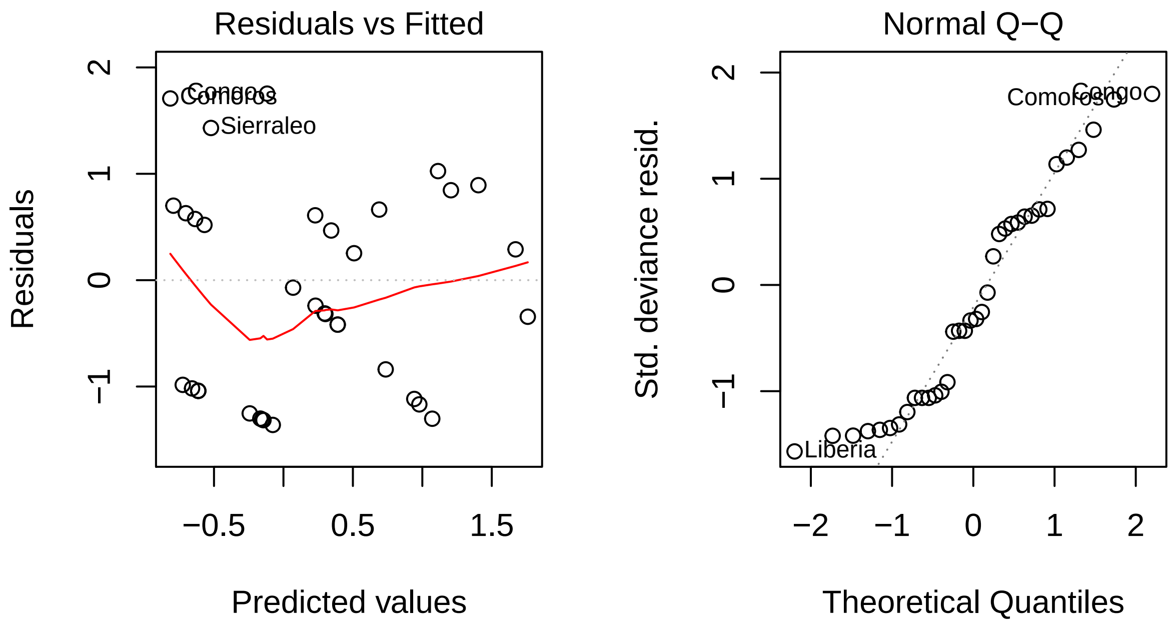
```
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.1165    0.51375  -0.227 0.820609
## oligarchy      0.0947    0.02318   4.085 0.000044
## pollib1       -0.6208    0.48753  -1.273 0.202919
## pollib2       -1.3104    0.48902  -2.680 0.007371
## parties        0.0257    0.00955   2.695 0.007036
## pctvote        0.0121    0.00907   1.329 0.183834
```

In this iteration, variable `pctvote` have the p-value > 0.05, thus we discard it for the next iteration.

```
#Final Model
model_poisson = glm(miltcoup~oligarchy+pollib+parties,family=poisson,data=africa)
summary(model_poisson)[[12]]
```

```
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.2080     0.4457   0.467 6.41e-01
## oligarchy      0.0915     0.0226   4.054 5.04e-05
## pollib1       -0.4954     0.4756  -1.042 2.98e-01
## pollib2       -1.1121     0.4595  -2.420 1.55e-02
## parties        0.0224     0.0091   2.458 1.40e-02
```

```
par(mfrow=c(1, 2))
plot(model_poisson, which = 1:2)
```



In the final model obtained using 'Step Down' approach, `oligarchy`, `parties`, and `pollib2` turned out to be most important variables with p-values less than 5%. Also, as per the QQ plot, distribution is evidently normaly distributed with few outliers.