assignment3

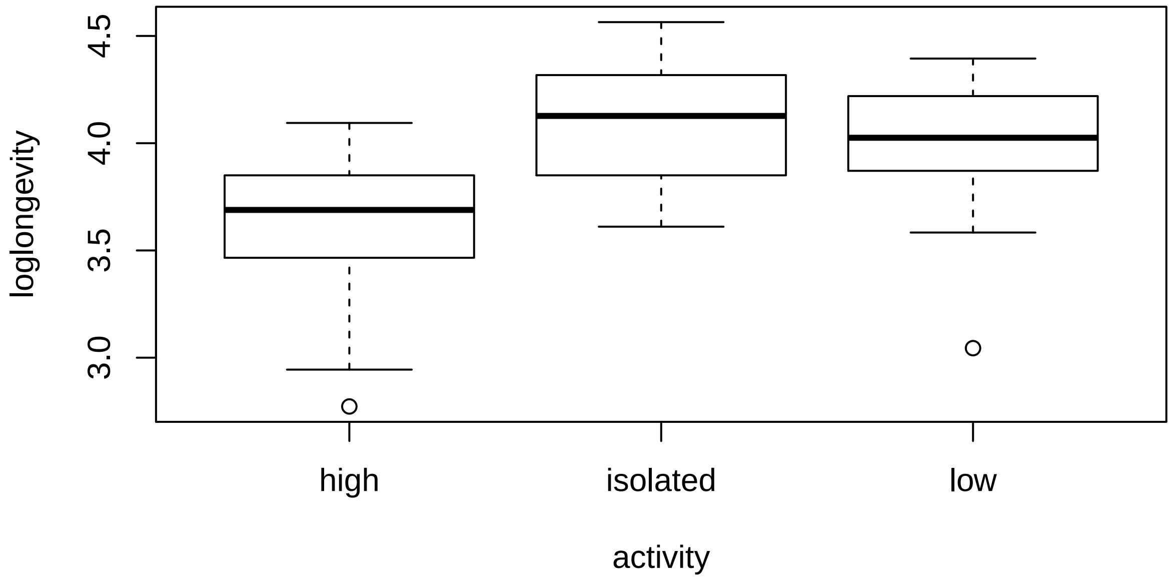
nihat uzunalioglu - 2660298, emiel kempen - 2640580, saurabh jain - 2666959

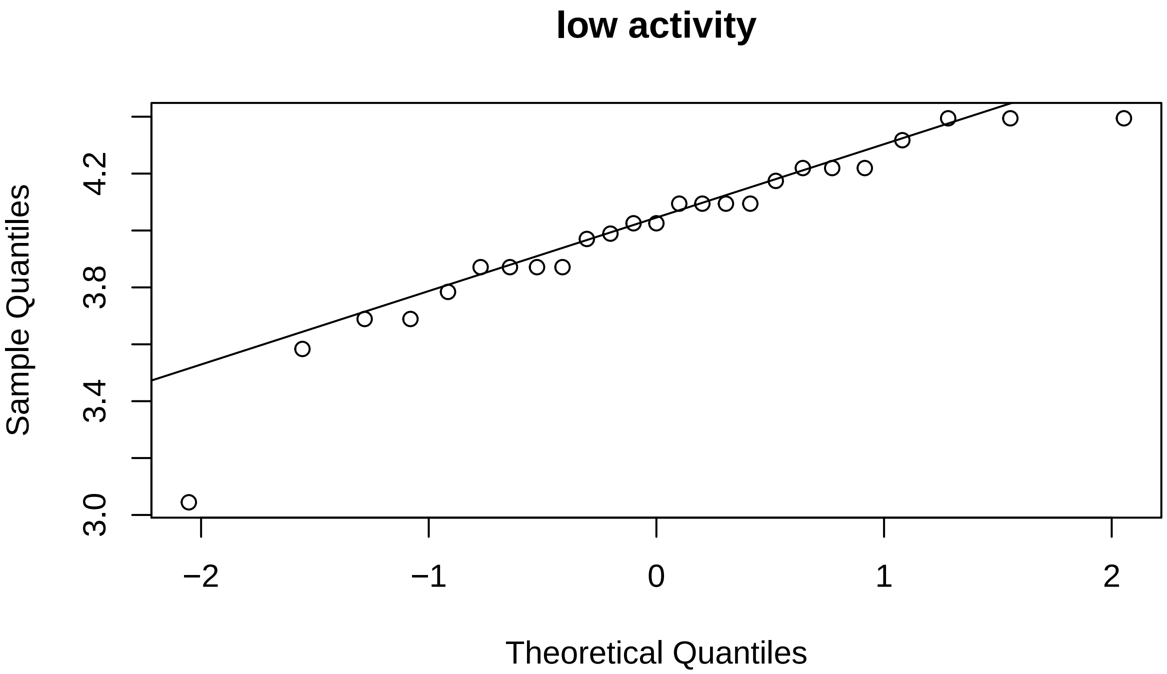
3/15/2020

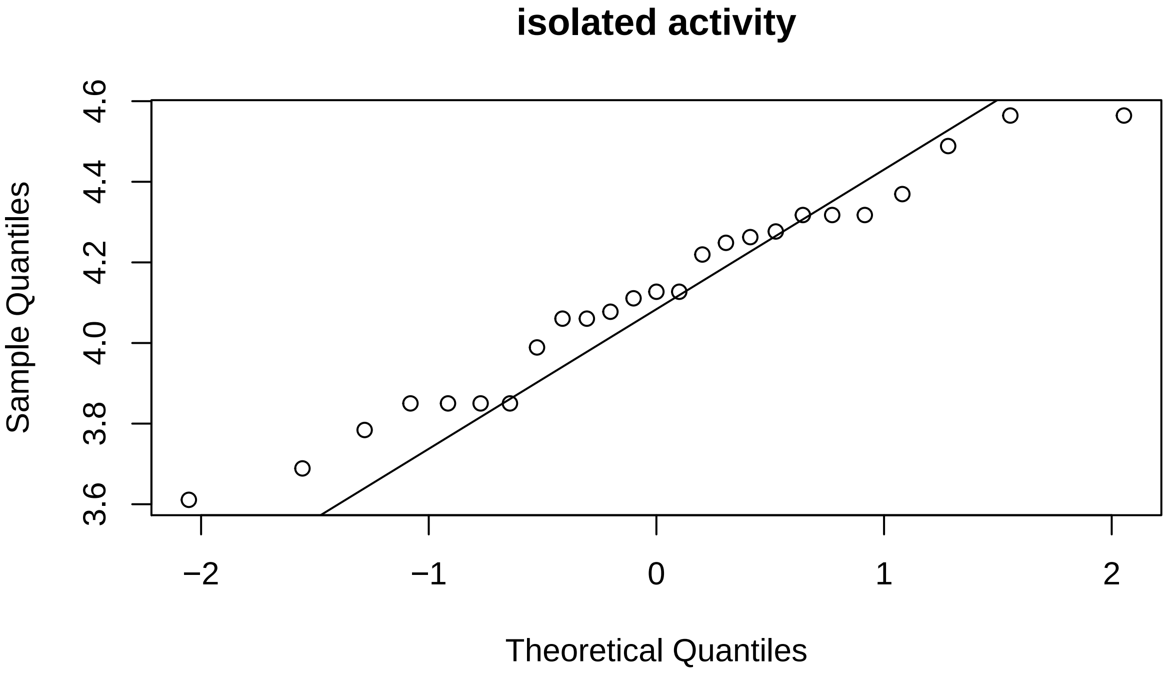
Exercise 1

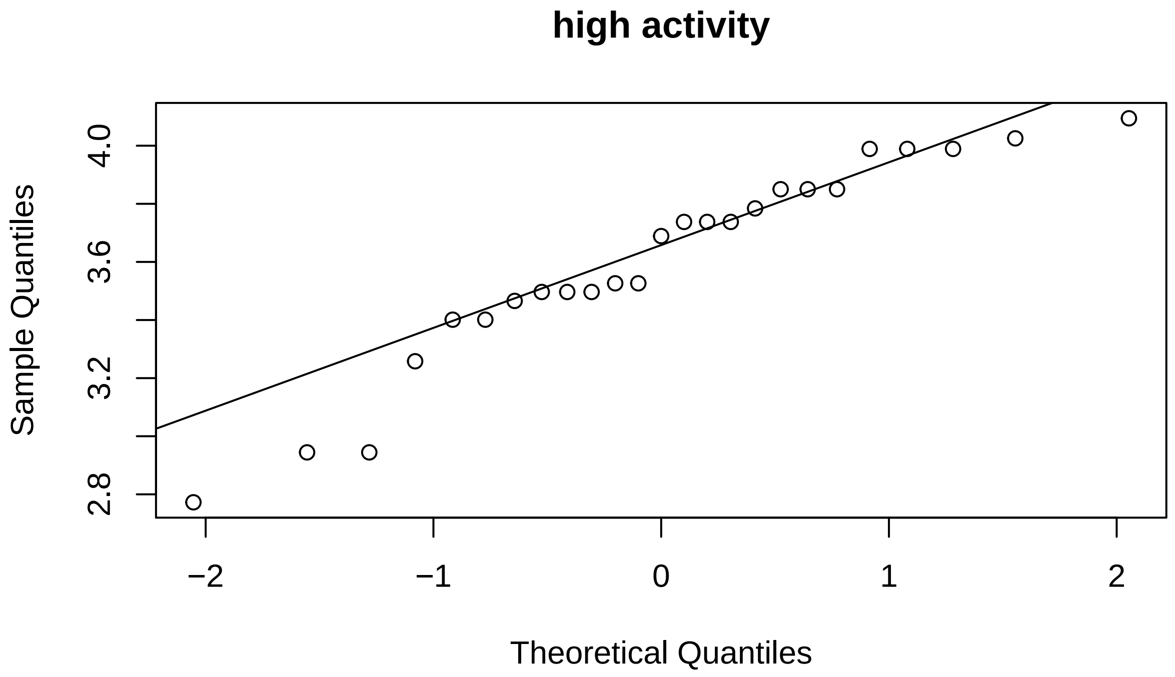
```
# Add loglongevity to the data-frame
# Use it as a response variable (Y)
fruitflies$loglongevity = log(fruitflies$longevity)
```

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

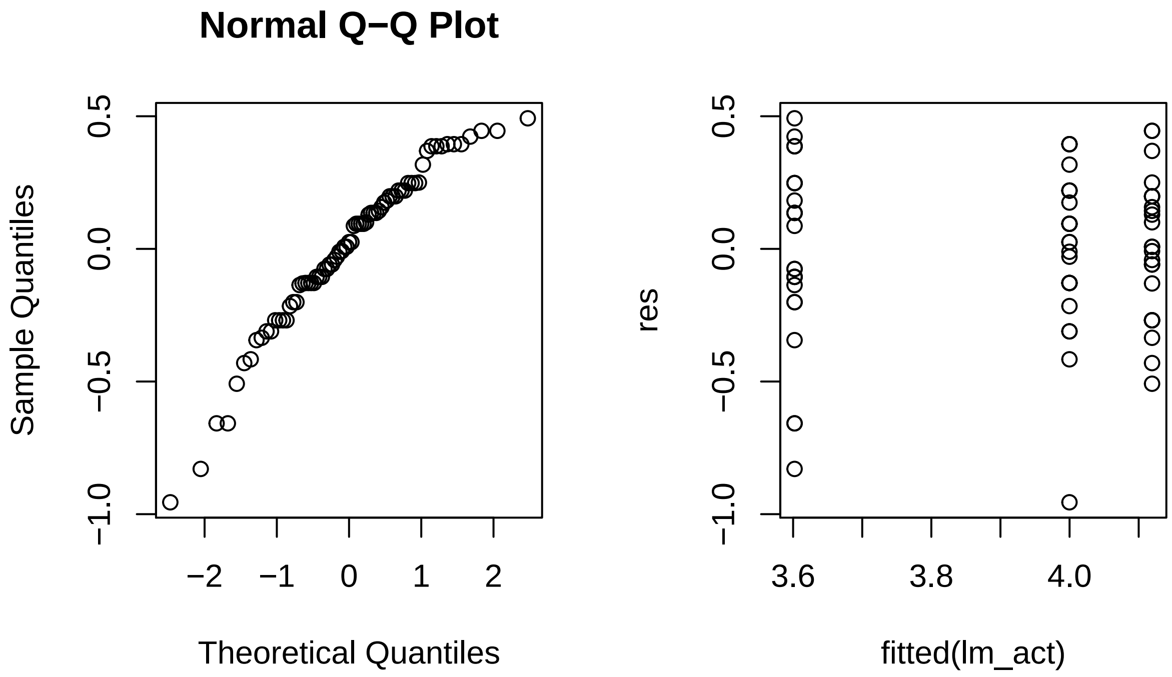


- From the boxplot, we can observe medians are not the same among activities and there is also an outlier both in high and low activity. From QQ-plots, we observed normality on low activity with 2 outliers, whereas isolated and high activities don't have normality.

```
## 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
```

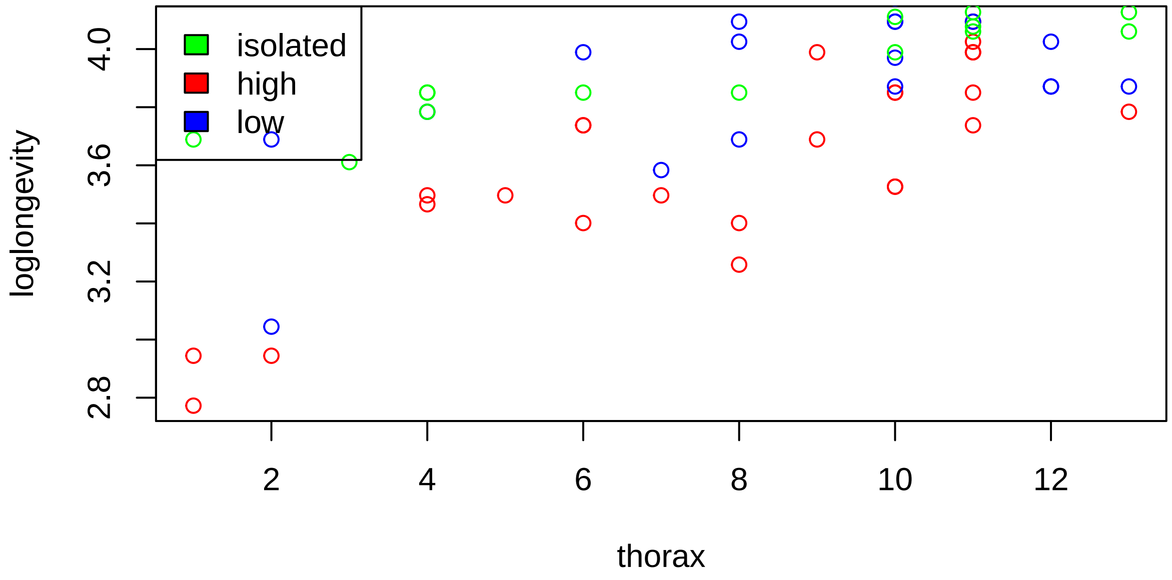
- According to the two-way ANOVA results, we observe that thorax (α_i) and activity (β_j) have both main effects on longevity in the additive model since we rejected the null hypothesis.
- From the summary coefficients, we observe that all activity factors increase longevity since all of them are positive.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

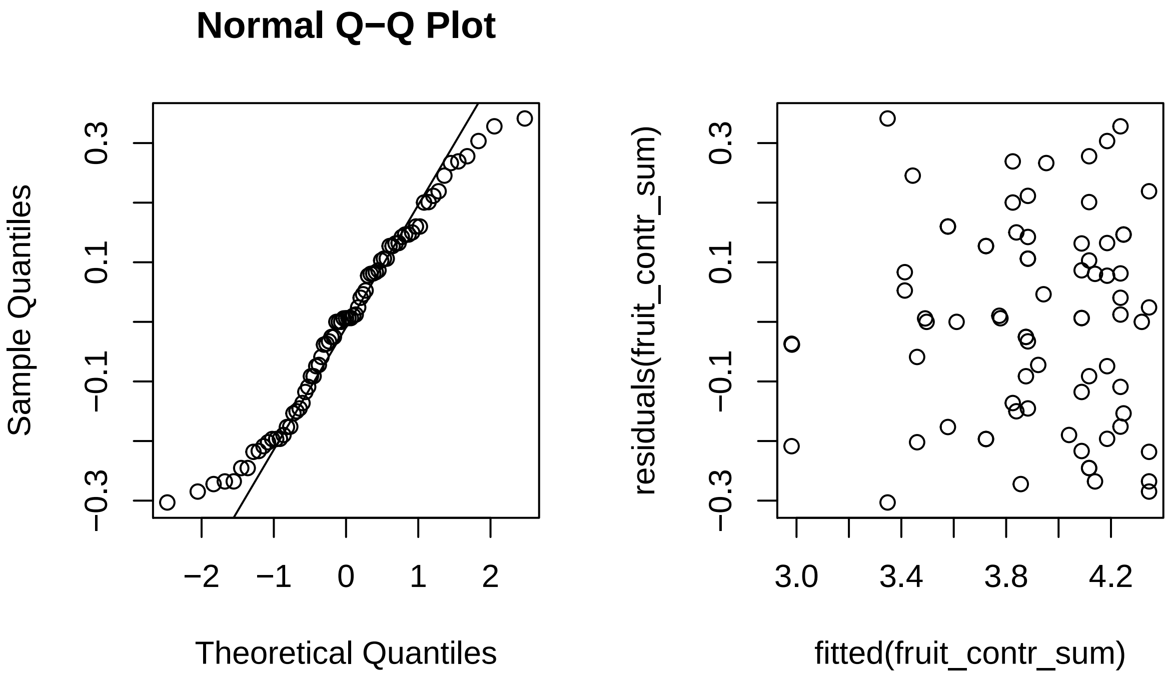
```
## $coefficients
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.0762    0.06758   45.52 2.82e-54
## thorax             0.0674    0.00693    9.72 1.10e-14
## activityisolated   0.4120    0.05832    7.07 8.92e-10
## activitylow        0.2871    0.05843    4.91 5.52e-06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

- 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



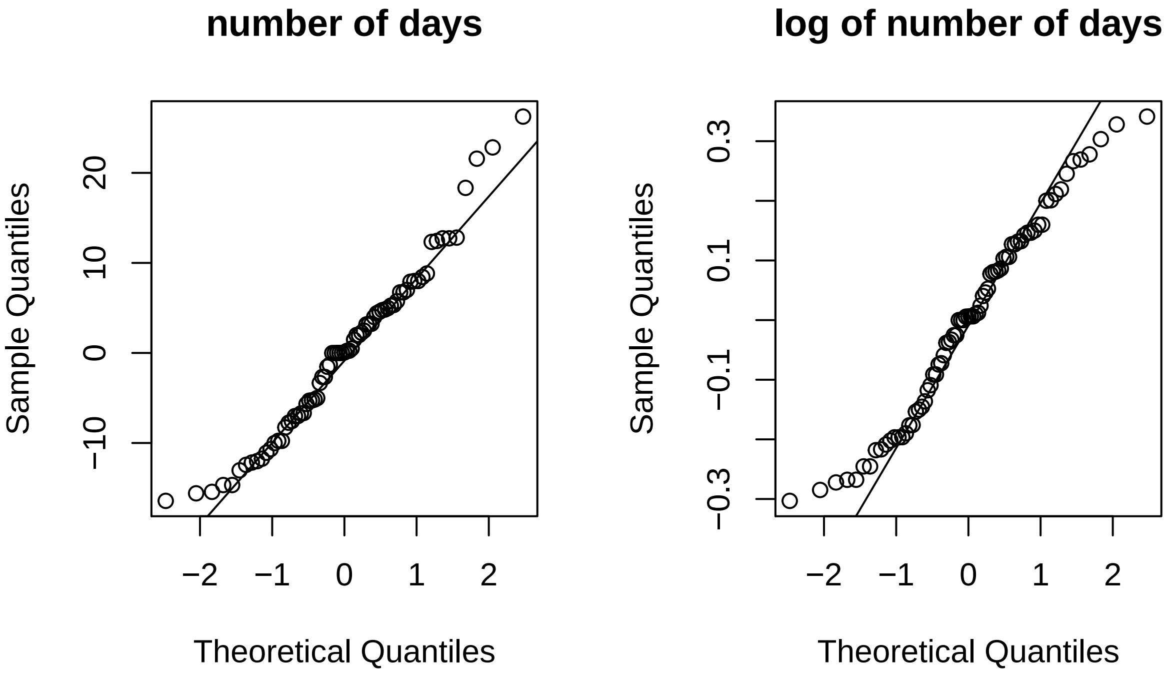
- The normality seems doubtful in QQ-plot and we did not observe heteroscedasticity in the scatter plot as they mostly are accumulated on the right side of the plot.

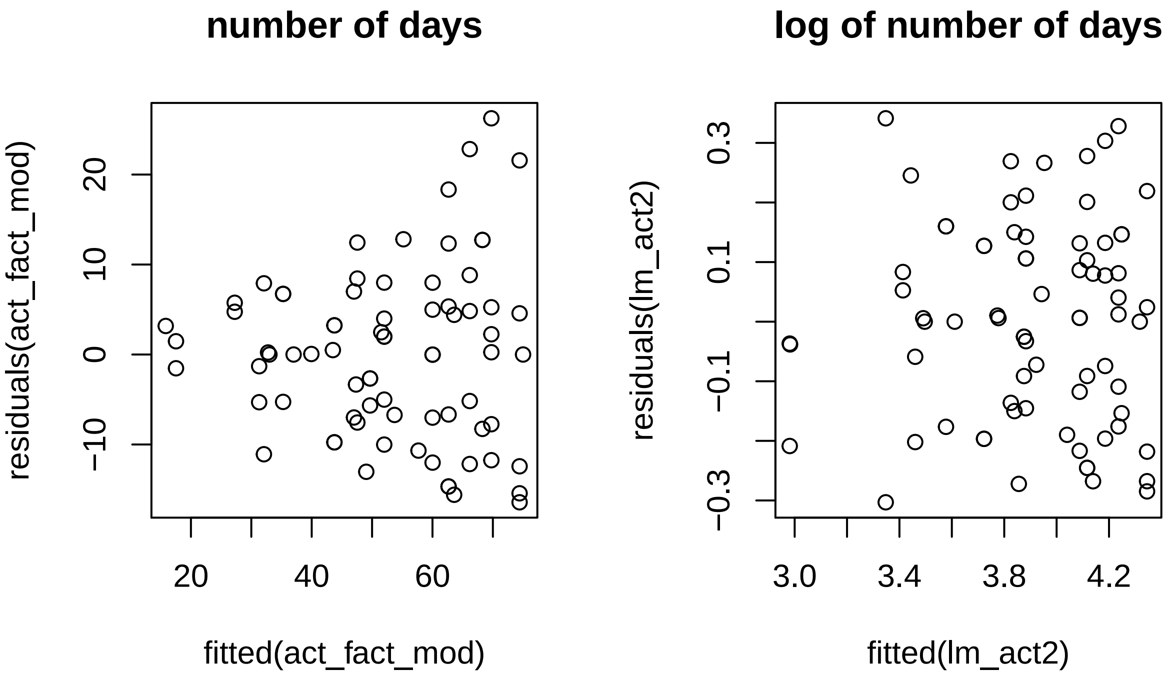
```
## Loading required package: zoo

##
## 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.