Assignment 3

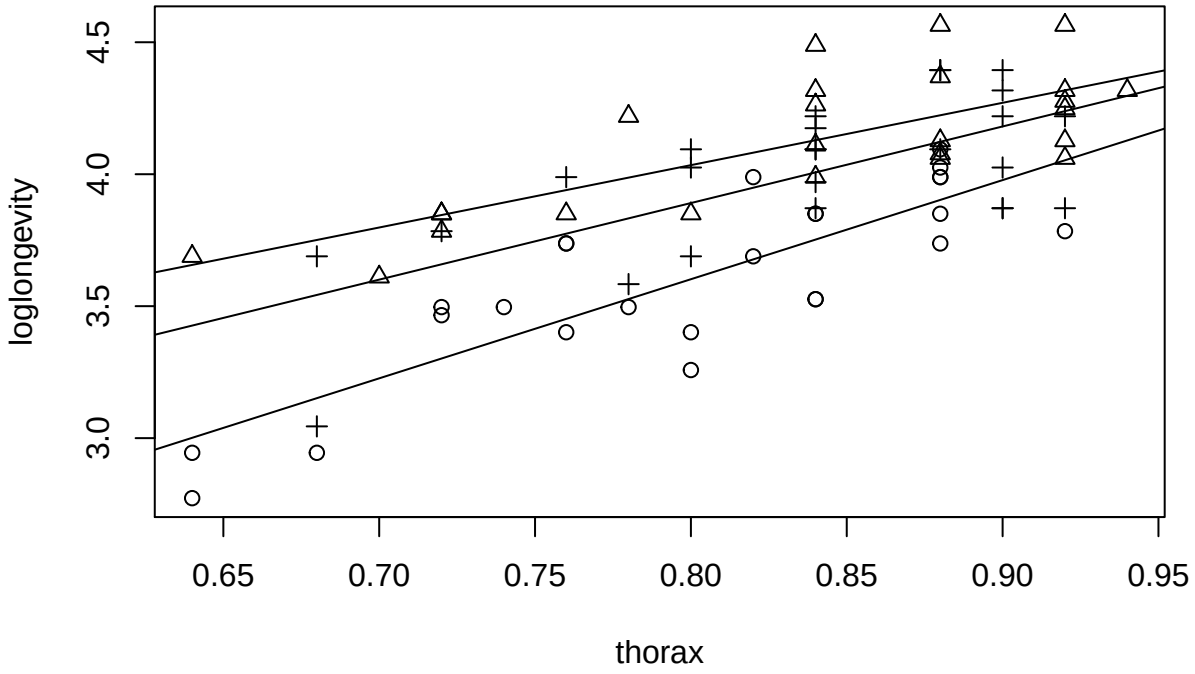
Kai Zhang, Sherida van den Bent, Chang Liu

2020/03/13

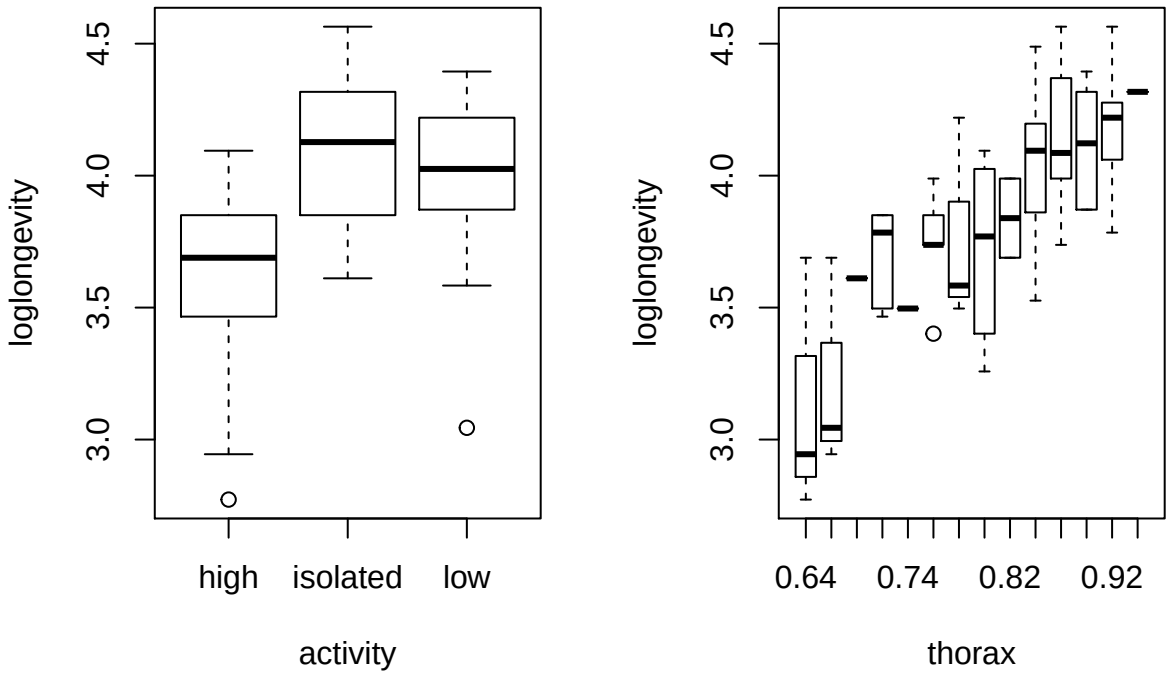
Exercise 1

a)

```
fruitflies = read.table('fruitflies.txt', header=TRUE)
fruitflies$loglongevity = log(fruitflies$longevity)
plot(loglongevity~thorax,pch=unclass(activity), data=fruitflies)
for (i in c('high', 'low', 'isolated')) abline(lm(loglongevity~thorax,data=fruitflies[fruitflies$activi
```



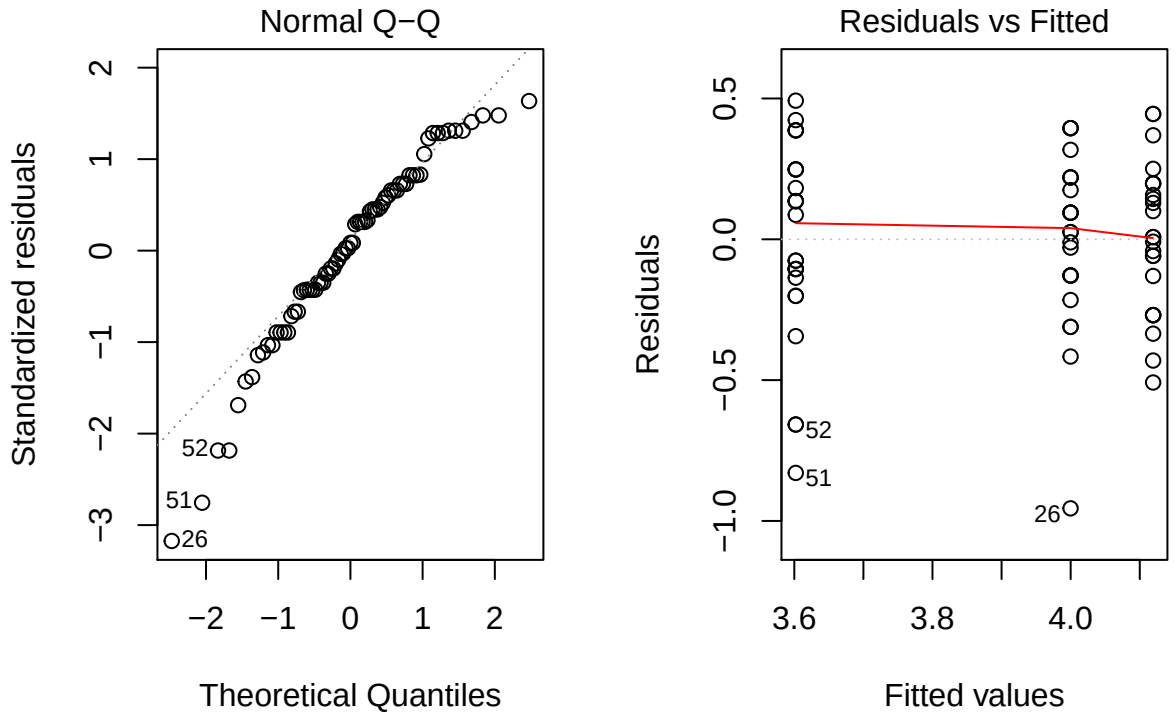

```
par(mfrow=c(1,2))
boxplot(loglongevity~activity, data=fruitflies); boxplot(loglongevity~thorax, data=fruitflies)
```




```
fruitfliesaov = lm(loglongevity~activity, data=fruitflies)
anova(fruitfliesaov); summary(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
##
## Call:
## lm(formula = loglongevity ~ activity, data = fruitflies)
##
## 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
```

```r
plot(fruitfliesaov, 2); plot(fruitfliesaov, 1)
```




```r
shapiro.test(residuals(object = fruitfliesaov))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(object = fruitfliesaov)
## W = 0.95431, p-value = 0.008652
```

From the first plot we can see that loglongevity clearly increases with thorax. Its dependence on activity is not so clear. From the two box plots, it is clear that both activity and thorax have effect on loglongevity. Then we use ANOVA to test $H_0 : \mu_{high} = \mu_{isolated} = \mu_{low}$. p-value = 1.798e-07 < 0.05. Conclusion we reject null hypothesis and sexual activity do influences longevity. From the summary, we get $\hat{\mu} = 3.60212$, $\hat{\alpha}_{isolated} = 0.51722$, $\hat{\alpha}_{low} = 0.39771$. So estimated loglongevities for high is 3.60212, for isolated is 3.60212+0.51722=4.11934 and for low is 3.60212+0.39771=3.99983. And estimated longevities for high is 36.67591, for isolated is 61.51863 and for low is 3.60212+0.39771=54.58887. We check the assumption of ANOVA and found Normal QQ-plot looks like has a curve, and Shapiro-Wilk normality test gives p-vlaue=0.008652, it means the conclusion is not reliable.

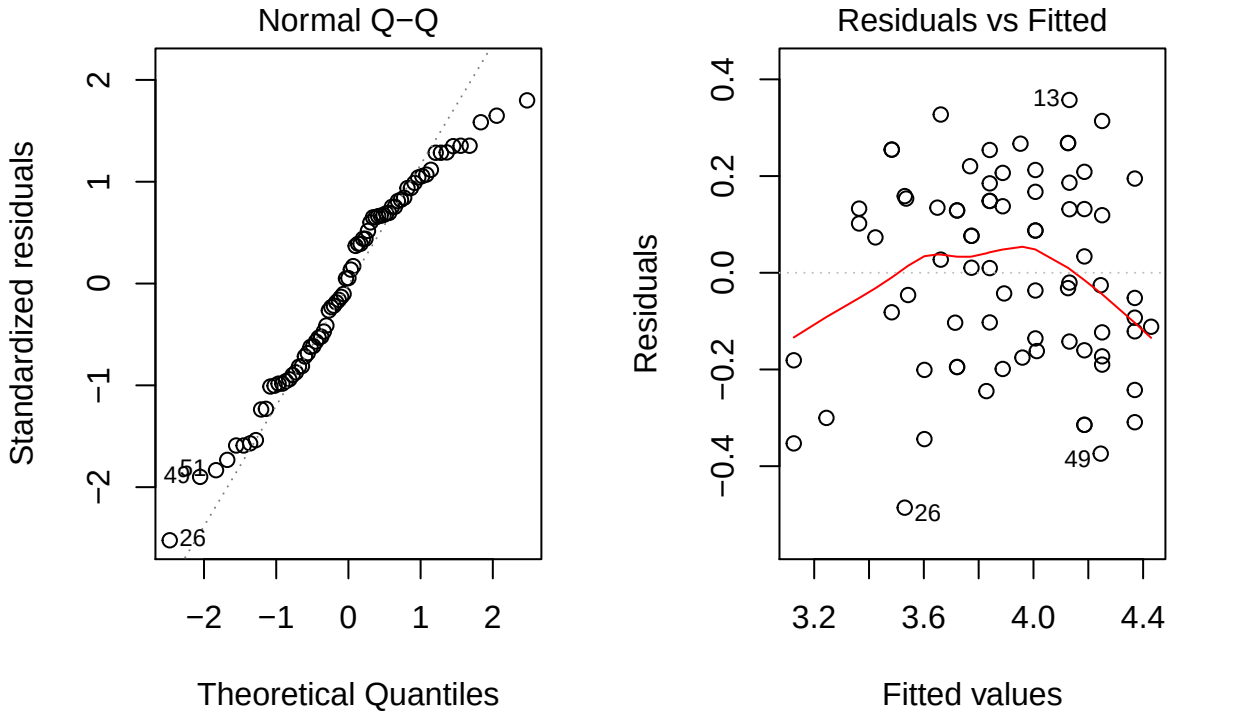
b)

```r
fruitfliesaov = lm(loglongevity~thorax+activity, data=fruitflies)
anova(fruitfliesaov); summary(fruitfliesaov)
```

```
## Analysis of Variance Table
##
## Response: loglongevity
##           Df Sum Sq Mean Sq F value Pr(>F)
## thorax     1 5.4322  5.4322 132.175 <2e-16 ***
## activity   2 2.1129  1.0565  25.705  4e-09 ***
## Residuals 71 2.9180  0.0411
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## 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)        1.21893    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
```

```
par(mfrow=c(1,2))
plot(fruitfliesaov, 2); plot(fruitfliesaov, 1)
```




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

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

```
shapiro.test(residuals(object = fruitfliesaov))
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  residuals(object = fruitfliesaov)
## W = 0.96838, p-value = 0.05748
```

We use ANCOVA to test $H_0 : \mu_{high} = \mu_{isolated} = \mu_{low}$. p-value = 4e-09 < 0.05. Conclusion we reject null hypothesis and sexual activity do influences longevity. From the summary, we get $\hat{\mu} = 1.21893$, $\hat{\alpha}_{isolated} = 0.40998$, $\hat{\alpha}_{low} = 0.28570$. So sexual activity decrease longevity We check the assumption of ANOVA and found Normal QQ-plot looks like has a curve, it means the conclusion is not reliable. The model is $\hat{Y}_i = \hat{\mu} + \hat{\alpha}_i + \hat{\beta} X_i$. The average thorax length is 0.8245333. The estimated loglongevity for a fly with average thorax length in isolated: $\hat{Y_{isolated}} = 1.21893 + 0.40998 + 2.97899 * 0.8245333 = 4.08518645537$. For low:$\hat{Y_{low}} = 1.21893 + 0.28570 + 2.97899 * 0.8245333 = 3.96090645537$. For High:$\hat{Y_{high}} = 1.21893 + 0 + 2.97899 * 0.8245333 = 3.67520645537$. So longevity for isolated is $e^{Y_{isolated}} = e^{4.09} = 59.45$. Longevity for low is $e^{Y_{low}} = e^{3.96} = 52.50$. Longevity for high is $e^{Y_{high}} = e^{3.68} = 39.46$. We check the assumption of ANCOVA. After creating a QQ-plot, we cannot tell whether it is a normal distribution because of some outliers. Therefore we also use the Shapiro-Wilk normality test. Based on the p-value($0.05748 > 0.05$), we could say that it is probably normally distributed. So the first assumption of normal distribution of the model residuals has been met. The fitted plot seems to indicate that the residuals and the fitted values are uncorrelated. We can rely on the ANCOVA test.

c)

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

```
##
## 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)                0.5978     0.4192   1.426   0.1584
## activityisolated           1.5465     0.5845   2.646   0.0101 *
## activitylow                0.9717     0.6423   1.513   0.1349
## thorax                     3.7554     0.5216   7.199 5.78e-10 ***
## activityisolated:thorax   -1.3929     0.7122  -1.956   0.0545 .
## activitylow:thorax        -0.8539     0.7794  -1.096   0.2771
## ---
## 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 plot in a), it is clear that the three lines would be parallel. P-value for $H_0 : \mu_{isolated} = \mu_{high}$ and $H_0 : \mu_{low} = \mu_{high}$ are 0.0545 and 0.2771. So we do not reject H_0. Conclusion: this dependence is similar under all three conditions of sexual activity.

d) I prefer the one with thorax length. Because the one without thorax length doesn't fit assumption of ANOVA.

e) From the plot from b), we can say the assumption of normal distribution of the model residuals and homogeneity of variance of the groups has been met.

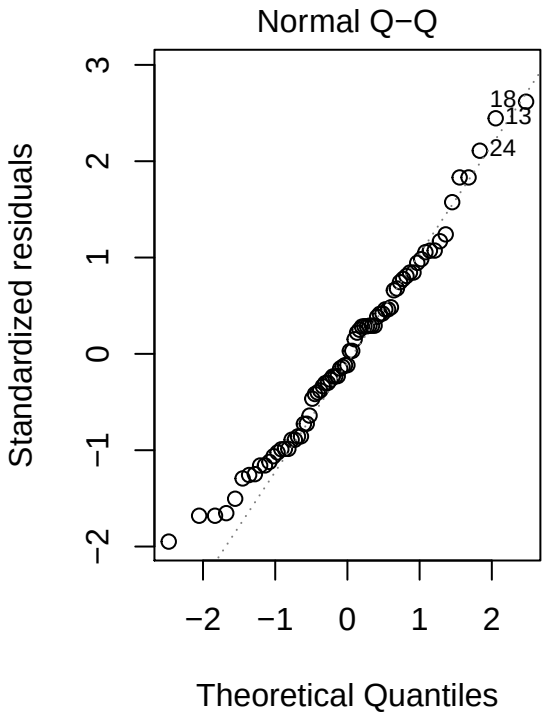
f)

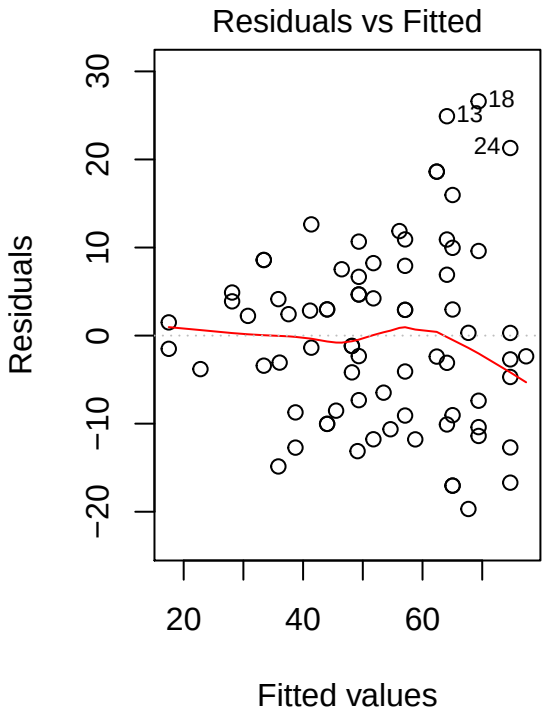
```
fruitfliesaov = lm(longevity~thorax+activity, data=fruitflies)
anova(fruitfliesaov); summary(fruitfliesaov)
```

```
## Analysis of Variance Table
##
## Response: longevity
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## thorax     1 10959.3 10959.3 101.409 2.557e-15 ***
## activity   2  4966.7  2483.4  22.979 2.016e-08 ***
## Residuals 71  7673.0   108.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Call:
## lm(formula = longevity ~ thorax + activity, data = fruitflies)
##
## 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)       -67.375     12.750  -5.284 1.33e-06 ***
## thorax            132.618     15.725   8.434 2.62e-12 ***
## activityisolated   20.066      2.994   6.701 4.13e-09 ***
## activitylow        13.054      2.999   4.352 4.43e-05 ***
## ---
## 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))
plot(fruitfliesaov, 2); plot(fruitfliesaov, 1)
```





From the fitted plot, we can see a certain pattern in residuals fitted plot. So the model with longevity instead of loglongevity is not reliable.