

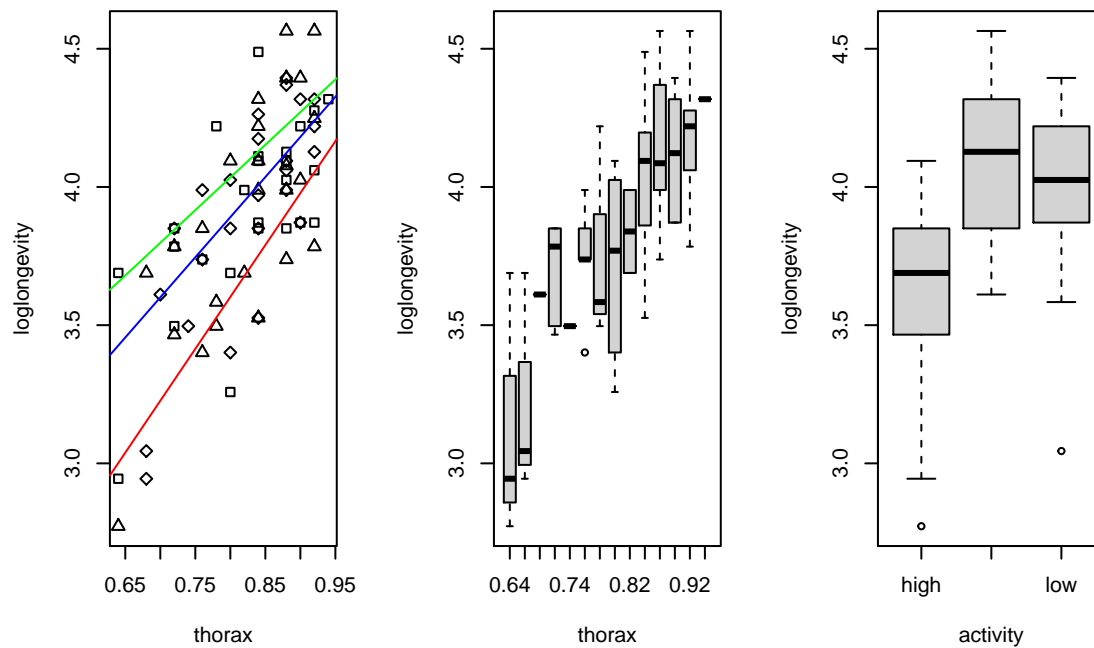
# A3E1

Xinyu Hu

2021/3/16

## Exercise 1.

a)



It is quite clear there exists a positive linear relationship between loglongevity and thorax. And the boxplots shows that loglongevity has influenced by thorax and activity.

```
ffaov = lm(loglongevity~activity, data=ffdata);anova(ffaov)
```

```
## 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
```

The anova test for testing  $H_0$  shows the  $p$ -value =  $1.798\text{e-}07 < 0.05$ , which concludes longevity is effected by activity.

```
summary(ffaov)
```

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

From above summary,  $\hat{\mu} = 3.60212$ ,  $\hat{\alpha}_{isolated}=0.51722$  and  $\hat{\alpha}_{low}=0.39771$ . Thus, estimated longevity of high  $e^{3.60212}=36.67$ , of isolated  $e^{3.60212+0.51722}=61.51$  and of low  $e^{3.60212+0.39771}=54.58$ .

b)

```
f1m = lm(loglongevity~thorax+activity,data=fdata);drop1(f1m,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
```

With Single term deletions,  $H_0$  is rejected because of  $p$ -value= $4.000\text{e-}09 < 0.05$ . It shows that activity still influences longevity.

```
summary(f1m)
```

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

The summary shows the  $\hat{\mu}=1.21893$ ,  $\hat{\alpha}_{isolated}=0.40998$  and  $\hat{\alpha}_{low}=0.28570$ . In this case, activity decreases longevity.

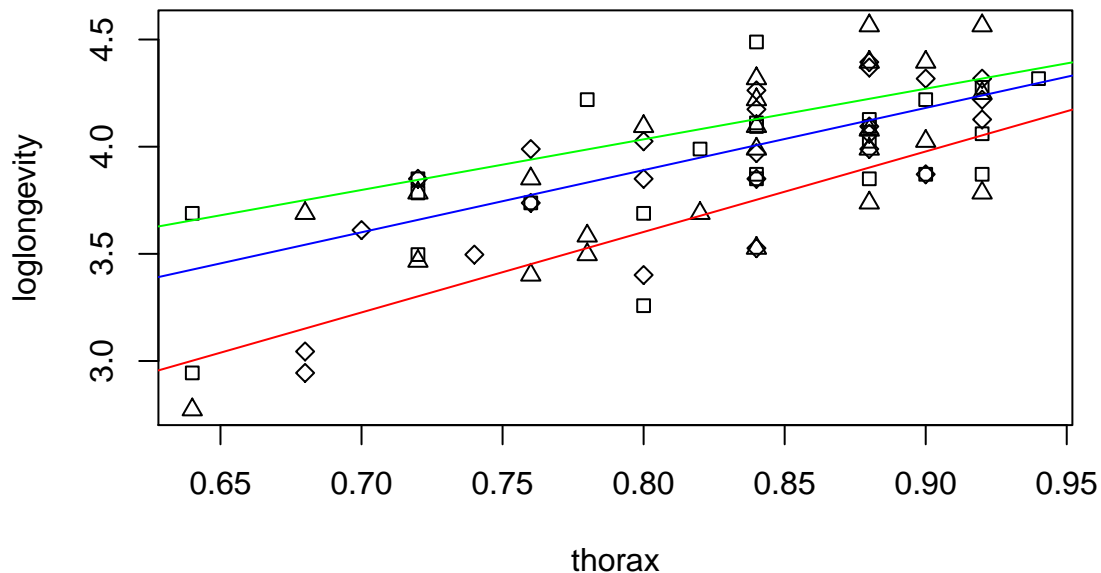
```
mean(ffdata$thorax)
```

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

The average thorax length = 0.8245333. Since the model is  $\hat{Y}_i = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}X_i$ ,  $\hat{Y}_{isolated} = 1.21893 + 0.40998 + 2.97899 * 0.8245333 = 4.085186455367$ ,  $\hat{Y}_{low} = 1.21893 + 0.28570 + 2.97899 * 0.8245333 = 3.960906455367$  and  $\hat{Y}_{high} = 1.21893 + 0 + 2.97899 * 0.8245333 = 3.675206455367$ . Additionally, the longevity of isolated  $e^{\hat{Y}_{isolated}} = e^{4.085186455367} = 59.45$ , of low  $e^{\hat{Y}_{low}} = e^{3.960906455367} = 52.50$  and of high  $e^{\hat{Y}_{high}} = e^{3.675206455367} = 39.45$ .

c)

### sexual activity and thorax length



```
fllmm=lm(loglongevity~activity*thorax,data=ffdata);summary(fllmm)
```

```
##
## Call:
```

```
## lm(formula = loglongevity ~ activity * thorax, data = ffddata)
##
## 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
```

The graph shows that each estimate for each group depending on the thorax length, and three lines are parallel. Additionally, from the summary, it shows that  $p$ -value for  $H_0 : \mu_{low} = \mu_{high}$  is  $0.2771 > 0.05$  and  $p$ -value for  $H_0 : \mu_{isolated} = \mu_{high}$  is  $0.0545 > 0.05$ . Therefore, we reject both of them, and this dependence is similar under all three conditions of sexual activity.

d)

```
summary(ffaov)$r.squared
```

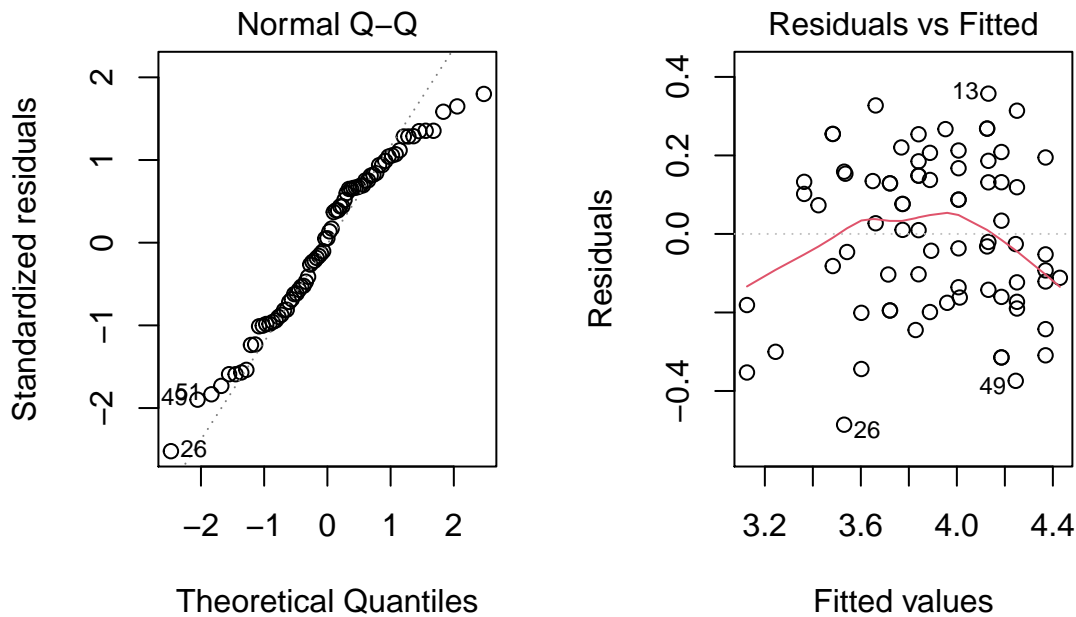
```
## [1] 0.3504222
```

```
summary(fflm)$r.squared
```

```
## [1] 0.721116
```

The analyses with thorax length is preferred, because the one with thorax length has a explained variance of 70.9%, but the one without thorax length has only 35%. None of them is wrong.

e)



```
mean(ffdata$thorax)
```

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

```
shapiro.test(residuals(fflm))
```

```
##
```

```
## Shapiro-Wilk normality test
```

```
##
```

```
## data: residuals(fflm)
```

```
## W = 0.96838, p-value = 0.05748
```

By checking QQ-plot and residuals vs. fitted plot, ANCOVA with thorax length and activity are independent to each other. In addition, we did not find certain patterns exist in the graph, even though they look normally distributed which is also checked by Shapiro-Wilk normality test.

f)

```
f flaov=lm(longevity~thorax+activity, data=ffdata)
anova(f flaov)
```

```
## 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 ***
```

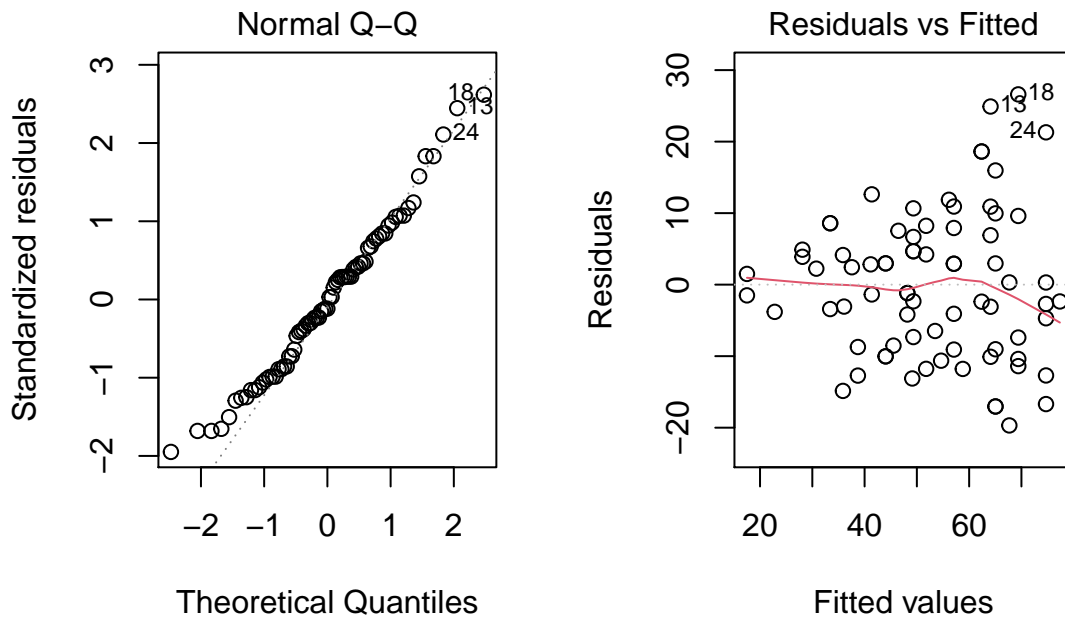
```
## Residuals 71  7673.0    108.1
```

```
## ---
```

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

Even without the logarithm, the ANCOVA shows very similar with the one with logarithm.

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



The graphs show that the residuals are normally distributed. And we can see a certain pattern existing in residuals fitted plot. This is also supported by Shapiro-Wilk normality test below.

```
shapiro.test(residuals(fflaov))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fflaov)
## W = 0.98091, p-value = 0.3176
```

However, the variance of residuals is getting larger as the the estimates gets larger. It does not reliable. Moreover, it is wise to use the model with logarithm.