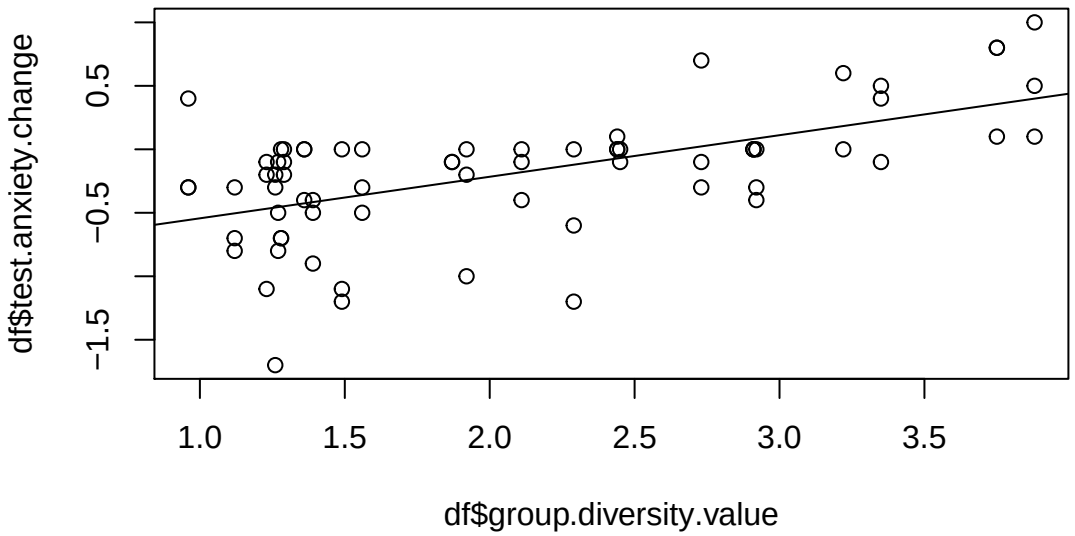
MAT325 Lab 4

Jordan Badstuebner

```
#############
##### 1 #####
#############
```

```
##
## Call:
## lm(formula = df$test.anxiety.change ~ df$group.diversity.value)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.24166 -0.24821  0.06875  0.25834  0.95655
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -0.87084    0.12500  -6.967 1.74e-09 ***
## df$group.diversity.value   0.32738    0.05566   5.882 1.42e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4061 on 67 degrees of freedom
## Multiple R-squared:  0.3405, Adjusted R-squared:  0.3307
## F-statistic: 34.59 on 1 and 67 DF,  p-value: 1.424e-07
```



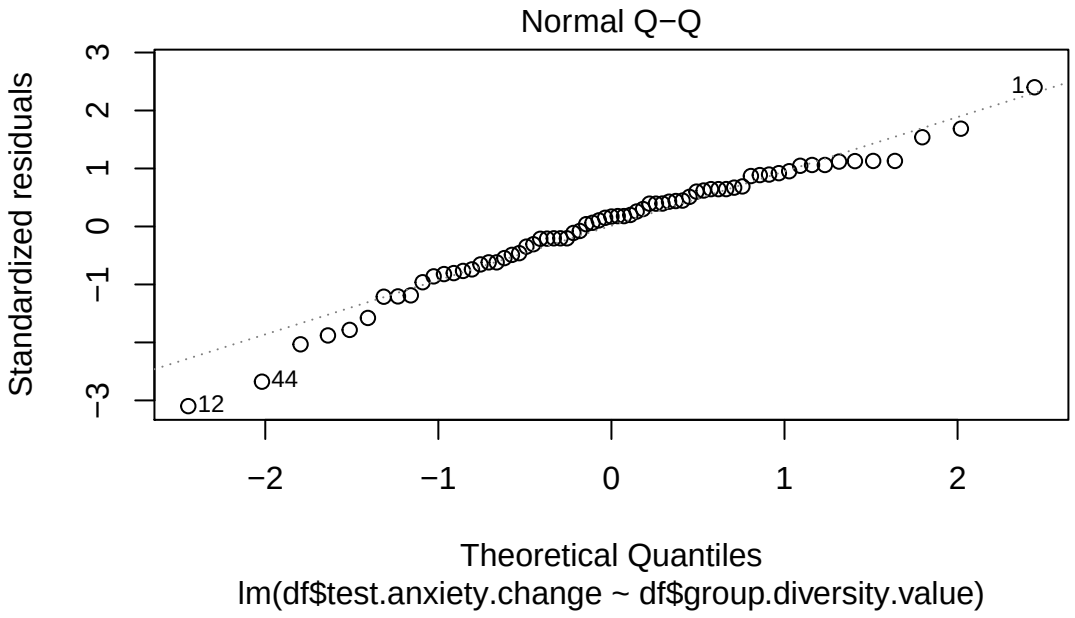
a) It appears the the majority of the data points are weighted toward the lower half of the x-axis. The trend of the data appears that it could possibly be sinusoidal.

b) The slope of the regression line is 0.32738, which predicts that as diversity value increases 1 unit, test anxiety increases 0.32738 units.

```
##                               2.5 %     97.5 %
## (Intercept)              -1.120345 -0.6213309
## df$group.diversity.value  0.216278  0.4384841
```

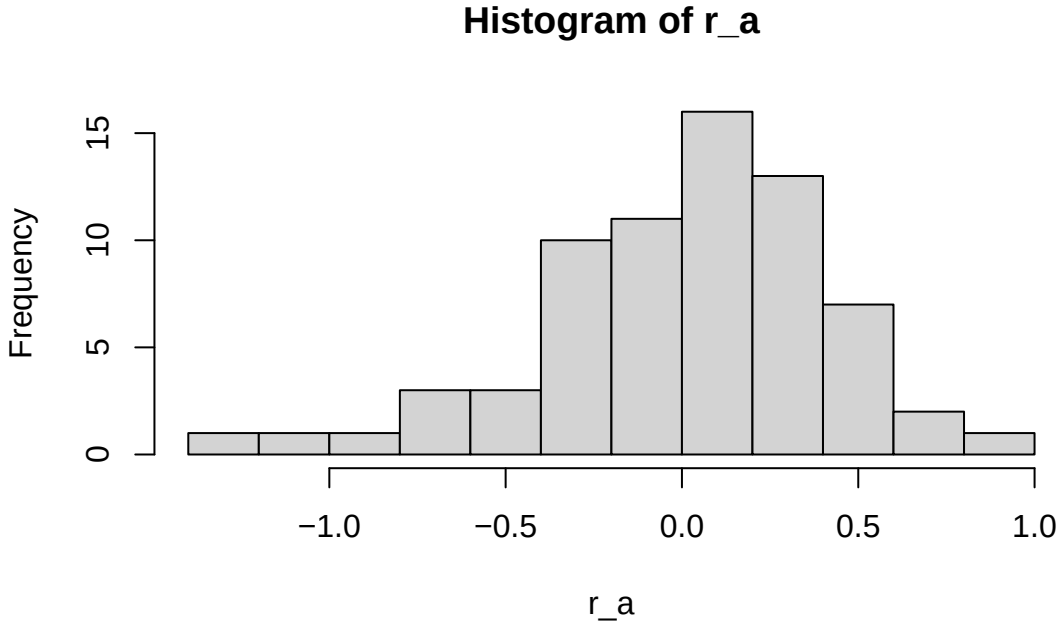
c) We are 95% Confident that the true slope lies between [0.216278, 0.4384841]

d) The Coefficient of determination of 0.3307 tells us that ~33% of the total variation is described by the variation in x (regression line).



e) The Normal Q_Q graph shows "heavy tails" vs Normal distribution. This means that the points fall along a line in the middle of the graph, but curve off in the extremeties. Heavy tails plots exhibiting this behavior ususally mean the data has more extreme values than would be expected if they truly came from a Normal distribution. This can be seen in the histrogram of the residuals below.

Source: https://data.library.virginia.edu/understanding-q-q-plots/



```r
# y_hat = b_0 + (b_1)*x  where b_0 is the intercept and b_1 is the slope

y_LB = -1.120345 + 0.216278*(4.5)
y_UB = -0.6213309 + 0.4384841*(4.5)

y_LB
```

```
## [1] -0.147094
```

```r
y_UB
```

```
## [1] 1.351848
```

f) We are 95% sure the change of test anxiety score for a person who is in a group with diversity value of 4.5 will be beween [-0.147094,1.351848]

```r
y_LB = -1.120345 + 0.216278*(3.3)
y_UB = -0.6213309 + 0.4384841*(3.3)

y_LB
```

```
## [1] -0.4066276
```

```r
y_UB
```

```
## [1] 0.8256666
```

g) We are 95% sure the change of test anxiety score for a person who is in a group with diversity value of 3.3 will be beween [-0.4066276,0.8256666]

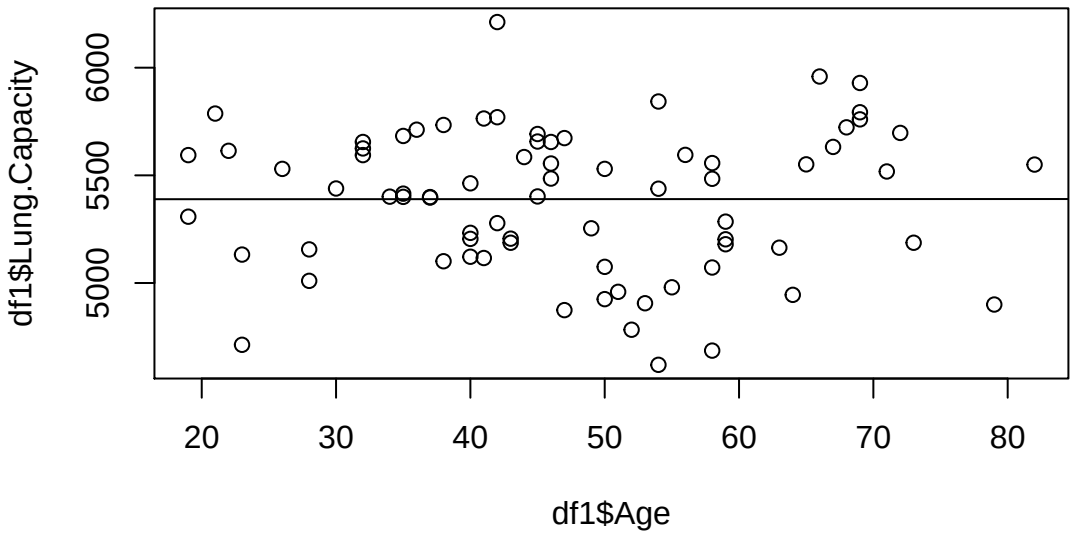
h) Observations 1, 12, and 44 could be an outliers.

```r
#############
##### 2 #####
#############
```

```
##
## Call:
## lm(formula = df1$Lung.Capacity ~ df1$Age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -769.63 -229.53   48.37  253.84  822.51
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.389e+03  1.281e+02  42.055   <2e-16 ***
## df1$Age     1.183e-02  2.590e+00   0.005    0.996
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 333.6 on 73 degrees of freedom
## Multiple R-squared:  2.859e-07,  Adjusted R-squared:  -0.0137
## F-statistic: 2.087e-05 on 1 and 73 DF,  p-value: 0.9964
```

a) The p-value is almost 1, which is highly insignificant, so there is almost definitely no linearly correlation between age and lung capacity. It can been seen in the plot below that the slope of the regression line is nearly 0.



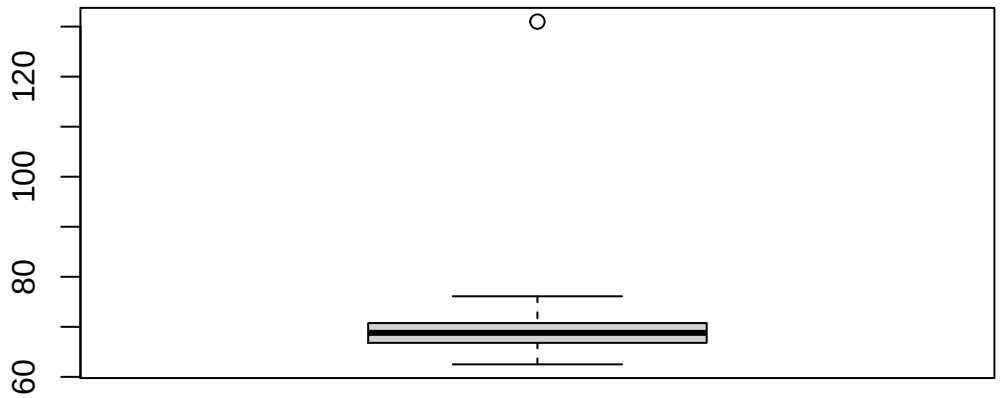

```
##
## Call:
## lm(formula = df1$Lung.Capacity ~ df1$Height)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -832.42 -199.32   40.15  240.27  711.67
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4215.224    325.345  12.956  < 2e-16 ***
## df1$Height    16.887      4.651   3.631 0.000521 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 307 on 73 degrees of freedom
## Multiple R-squared:  0.153,  Adjusted R-squared:  0.1414
## F-statistic: 13.18 on 1 and 73 DF,  p-value: 0.000521
```

b) There seems to be a linear correlation between height and lung capacity because the p-value of .000521 shows that the slope is sigficant. The R-squared value is 0.1414 though which tells our that our model

isn't great. The slope is 16.887 which means that for every increase in unit of height, lung capacity increases 16.887 units.

```
##                    2.5 %     97.5 %
## (Intercept) 3566.812646 4863.63587
## df1$Height     7.618242   26.15577
```

It can be seen that we are 95% confident that the slope lies between [7.618242, 26.15577] units.



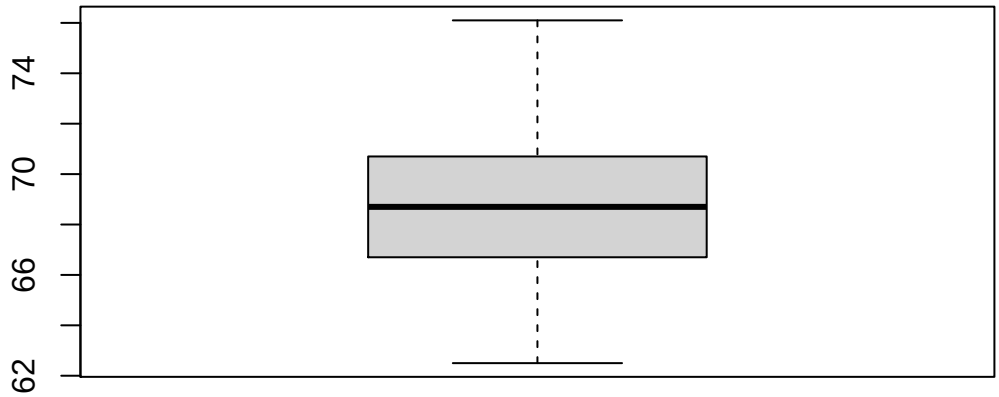

```
## [1] 131
```

The height of 131.0 units appears to be an outlier. The units of measurement are not provided in the data, but if this is 131 cm the person would be ~1.5 meters tall and if this is 130 in. then the person is over 9 ft. tall. I think it is safe to assume that this height was most likely an error made while recording the data.

```
# Find rows containing the outliers
df1[which(df1$Height %in% outliers),]
```

```
##    Height Exercise Age Lung.Capacity
## 40    131       21  56          5595
```

```
# Remove the rows containing the outliers
df1_HOR = df1[-which(df1$Height %in% outliers),]
boxplot(df1_HOR$Height)
```



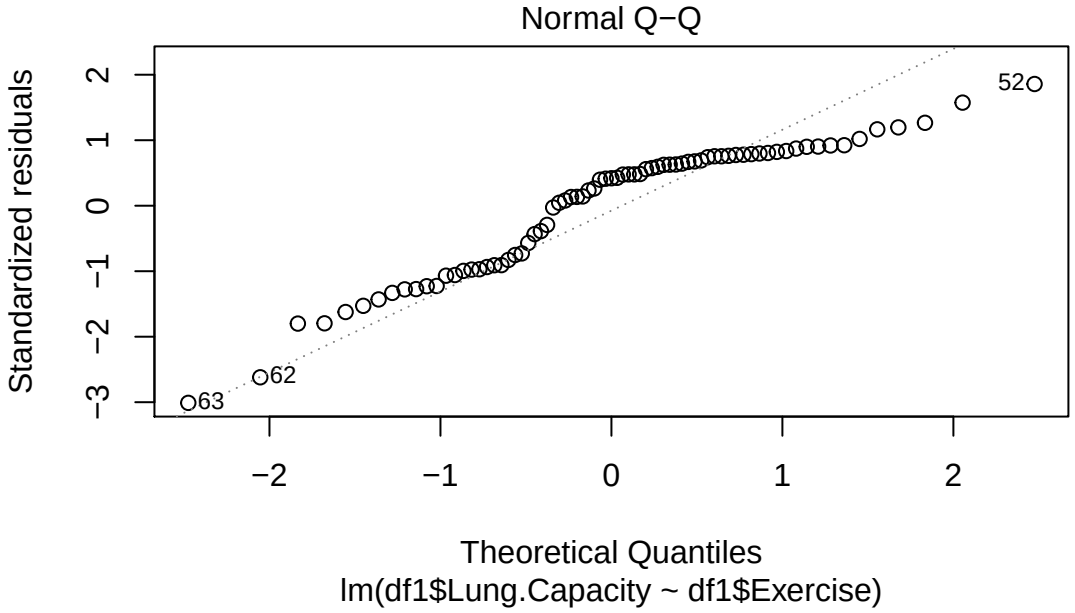

```
##
## Call:
## lm(formula = df1_HOR$Lung.Capacity ~ df1_HOR$Height)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -294.29  -81.28   -1.71   72.92  389.85
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -2487.751    371.167  -6.703 3.89e-09 ***
## df1_HOR$Height   114.606      5.398  21.232  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 124.3 on 72 degrees of freedom
## Multiple R-squared:  0.8623, Adjusted R-squared:  0.8604
## F-statistic: 450.8 on 1 and 72 DF,  p-value: < 2.2e-16
```

As we can see above, removing the outlier in height has made substantial difference. The P-value is much lower, the F-stat is far higher making the slope of regression (which is now far steeper) much more significant. There now appears to be a much stronger linear correlation, supported by a R-squared value of now 0.8604!

```
##
## Call:
## lm(formula = df1$Lung.Capacity ~ df1$Exercise)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -880.4 -267.9  123.4  223.8  506.6
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4895.422    118.616  41.271  < 2e-16 ***
## df1$Exercise   22.365      5.139   4.352 4.32e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 297.3 on 73 degrees of freedom
## Multiple R-squared:  0.206,  Adjusted R-squared:  0.1951
## F-statistic: 18.94 on 1 and 73 DF,  p-value: 4.317e-05
```

c) There seems to be a linear correlation between exercise and lung capacity. The p-value is 4.32e-05 which is very low showing that our regression slope of 22.365 is quite significant. Our R-squared value is 0.1951 which tells us that ~20% of the total variation is described by the variation in the regression line. This means that our model isn't great.



It seems to me that the residuals are a negative skew or possibly symmetric with fat tails. A negative skew in the data distribution would make more sense to me, inferring that those with the highest exercise ratings possess a higher lung capacity, but those who are average or low exercise ratings possess lower and less varied lung capacity.