Griffin Statistics

Jonah Zembower, Stanislav Chernyshov, Aidan Owens, Christian Zilli, Cole Kuczynski

May 2, 2023

1. We are first analyzing the relationship between load in Newtons on a sled and the time it takes for the sled to reach the top of a ramp. Particularly as load increases, will there be an effect on the time? First, we will create a regression model with this in mind. From the model below, we see that the p-values are relatively low but not below 0.05 for a 95% confidence interval. Furthermore, the R^2 and F-statistic values are very low at 0.1777 and 3.457 respectively, which can doubt the predictiveness of the model.

```
Residuals:
    Min      1Q  Median      3Q     Max
-121.49  -32.75  -19.15    8.97  433.47

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   985.03     501.50   1.964   0.0671 .
Load          -11.14       5.99  -1.859   0.0815 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 123.7 on 16 degrees of freedom
Multiple R-squared:  0.1777,    Adjusted R-squared:  0.1263
F-statistic: 3.457 on 1 and 16 DF,  p-value: 0.08148
```

While trying to assess why this model would be this poor, we examined the particular data set and noticed below, as highlighted, that many extreme values can be removed

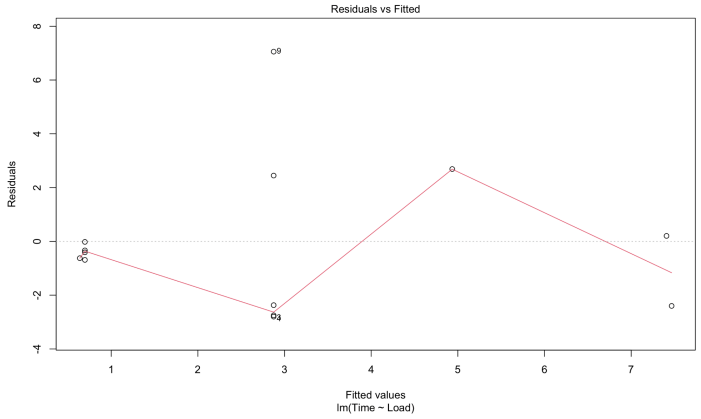
Load	Time
77.7	5.067
77.8	552.056
77.9	127.809
77.8	7.611
85.5	0.124
85.5	0.077
89.2	0.008
89.3	0.013
73.1	49.439
85.5	0.503
89.2	0.362
85.5	9.930
89.2	0.677
85.5	5.322
89.2	0.289
82.3	53.079
82.0	7.625
82.3	155.299

.

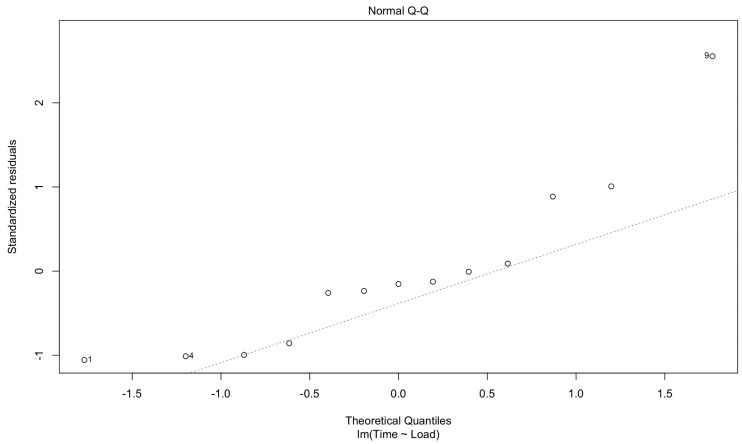
After removing the values, here is the new regression model. We see that this model has a decreased p-value of less than 0.05. Also, the R^2 and F-statistic values have increased. We can assume that removing the values above is more realistic in predicting the model.

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.7978 -2.3718 -0.4076  0.2031  7.0552

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  53.2096    17.3298   3.070   0.0107 *
Load         -0.5887     0.2025  -2.907   0.0143 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.876 on 11 degrees of freedom
Multiple R-squared:  0.4344,    Adjusted R-squared:  0.383
F-statistic: 8.448 on 1 and 11 DF,  p-value: 0.01428
```

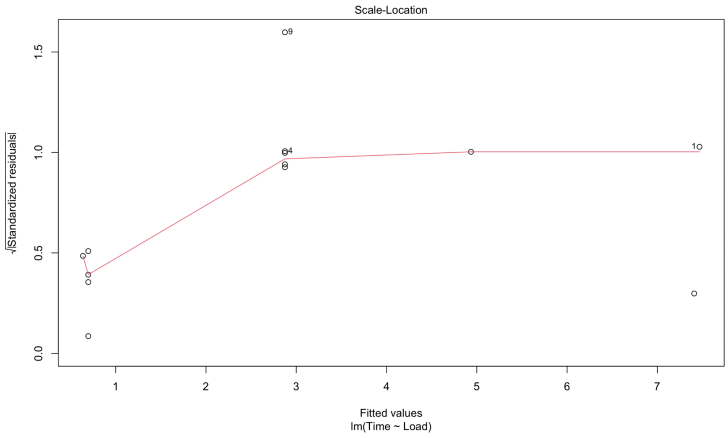
Now we would be checking the autocorrelation to see how explanatory variables are related to each other. However, there is only one variable, therefore, it is unnecessary. Next, we would check through the subset selection method, but there is only one explanatory variable, so we will assume that the assumptions are followed. Since our regression model has stayed the same, we will look at the various plots to check for linearity, normality, homoscedasticity, and leverage. For the first assumption of linearity, we do not accept the assumption because the red line is not horizontal, and the data values showcase why that is occurring with differing residuals.



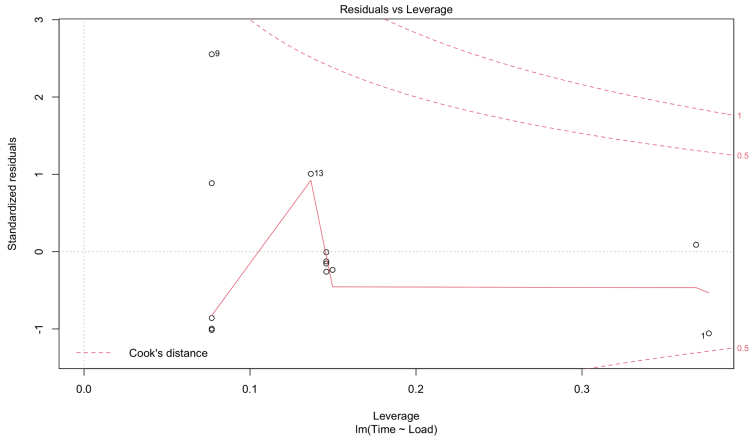
For the second assumption of normality, we reject the assumption since the values don't appear to follow the line. There is one specific data value in the top right that is very different.



For the third assumption of homoscedasticity, we reject the assumption because the red line is once again not horizontal as we would expect. This is due to some smaller standardized residual values at the start.



Regarding the last assumption of leverage, we relatively accept the assumption, as there appear to be no influential points affecting the model. There are no values past Cook's distance.



Overall, we conclude that this model is not very effective in its predictability of the time that it takes for a sled to reach the top of a ramp based on the load in Newtons. This is due to the low multiple R^2 value of 0.4344. Furthermore, since only one assumption was accepted, the model

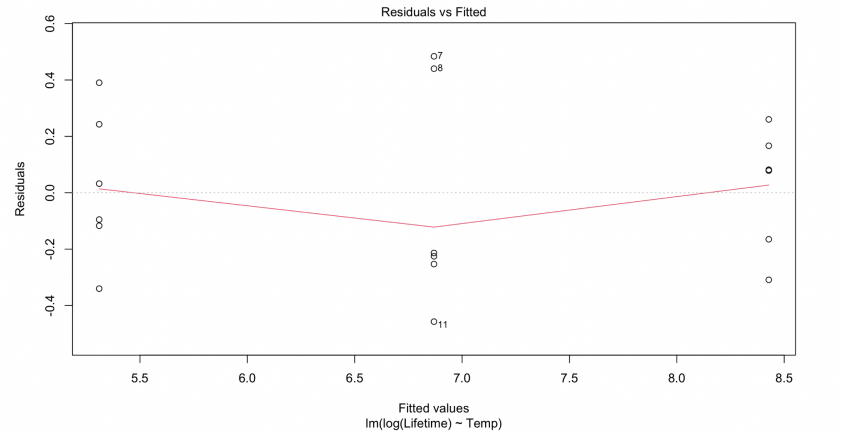
does not appear to follow what we would hope for in a good model. Therefore, we would strongly advise against the use of this model in the particular context given for the problem.

2. For the second problem, we are assessing how heat temperature affects the lifetime of a particular bacteria. We have three temperature zones of 200, 220, and 240 that will be analyzed as the explanatory variable temperature with the respective lifetime of the bacteria. First, we will develop the logarithmic linear regression model below due to this being expressed over a lifetime. From this model, we can assess the p-value to be very low, there is a high multiple R^2 value of 0.9548, and the F-Statistic is very high as well at 338.3. This model appears to have great potential in the predictability of the lifetime of the bacteria.

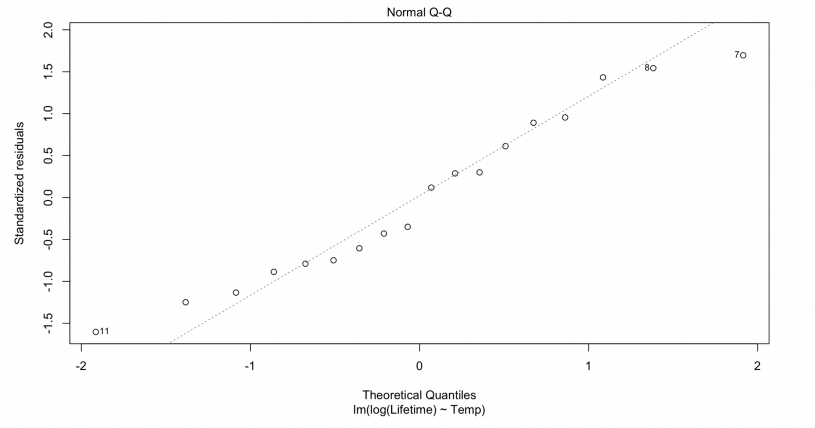
```
Residuals:
     Min       1Q   Median       3Q      Max
-0.45732 -0.22244 -0.03149  0.22381  0.48394

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.018336   0.934966   25.69 1.96e-14 ***
Temp        -0.077951   0.004238  -18.39 3.47e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2936 on 16 degrees of freedom
Multiple R-squared:  0.9548,    Adjusted R-squared:  0.952
F-statistic: 338.3 on 1 and 16 DF,  p-value: 3.469e-12
```

Next, we will not need to check autocorrelation or any specific selection methods due to one variable. Now we will check the remaining assumptions:

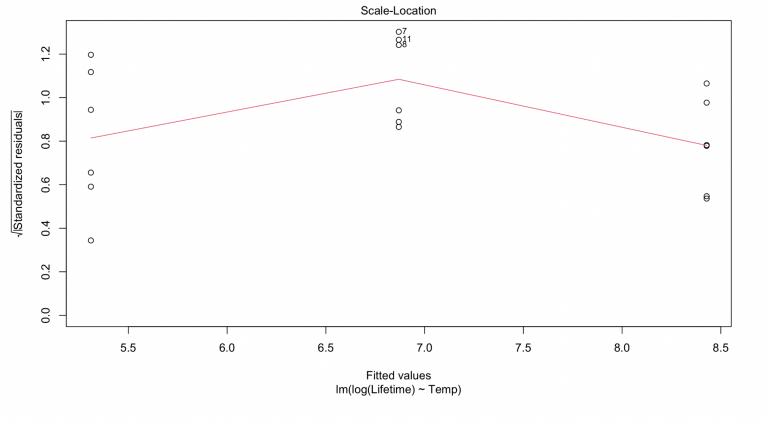
For the first assumption of linearity, we accept the assumption because there is a relatively horizontal line. The residuals appear to balance out for this model.



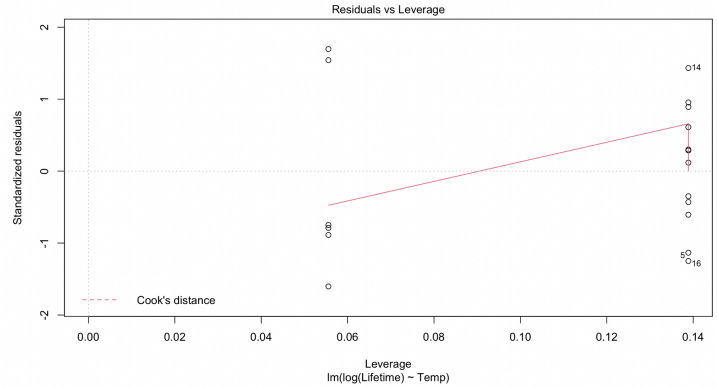
For the second assumption of normality, we accept the assumption because the values relatively follow a line. There is slight oscillation, but it still generally follows the line.



For the third assumption of homoscedasticity, we reject the assumption because the values don't follow a horizontal line. This is because the residuals in the middle are higher than the others.



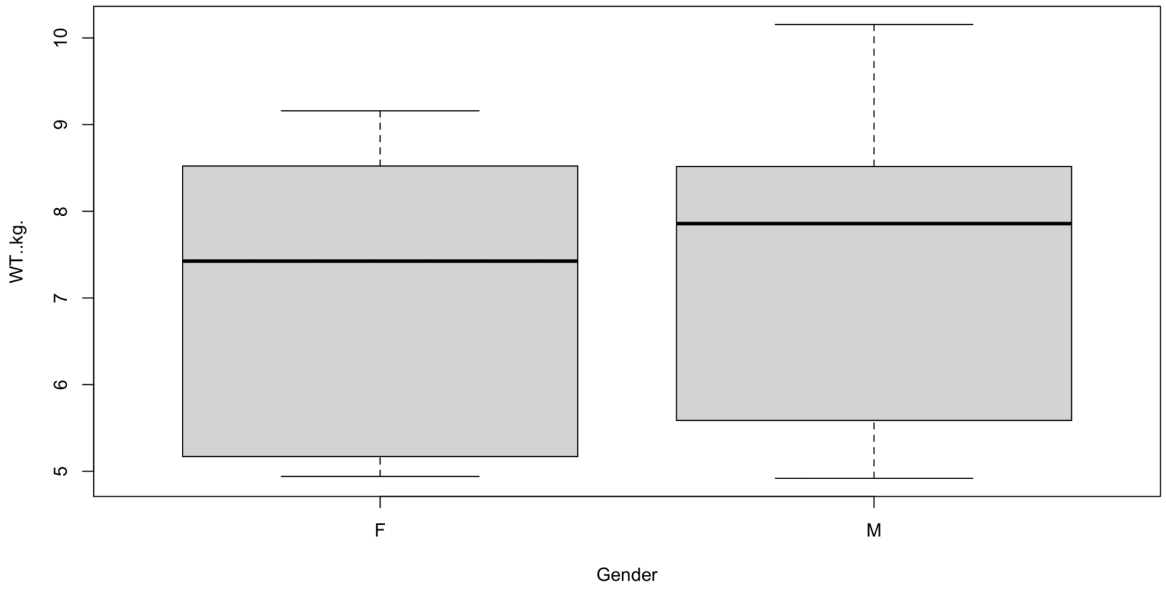
For the fourth assumption, there were no influential points. Therefore, we will accept the assumption that there are no influential points. There were no values past Cook's distance.



Overall, from the statistics we have found and that three assumptions were accepted, we can relatively believe that the logarithmic linear regression model effectively predicts the lifetime of bacteria when heated at temperature zones 200, 220, and 240.

3. For this part, we are analyzing the Caiman in the Amazon Basin. There are many variables we are assessing, such as weight, gender (male or female), and genetic marker type (Type 1 or Type 2).

First, we are assessing the difference between the average weight in kg of the Caimans based on gender. To perform this, we are implementing a box plot that showcases male and female weights respectively. From this graph, we can assess that the male Caiman are relatively heavier in weight with a higher median, quartiles, and the highest value of weight rather than females.



Next, we are checking the variance between the two populations of male and female Caiman using the F-test to compare the two variances with a null hypothesis that there is no difference and if, statistically significant, there is a difference. Due to the F-test having a p-value of 0.7851 that is not statistically significant, we accept that the two populations of male and female Caiman have similar variances.

```
       F        M
2.441406 2.632952

Rcmdr>  var.test(WT..kg. ~ Gender, alternative='two.sided', conf.level=.95,
Rcmdr+    data=Problem_3)

        F test to compare two variances

data:  WT..kg. by Gender
F = 0.92725, num df = 54, denom df = 46, p-value = 0.7851
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5249078 1.6171728
sample estimates:
ratio of variances
         0.9272506
```

Now we are checking for the range we feel 95% confident that the proportion of Caiman with Type II will fall in such an interval. We first find the sample percentage of Type II to be 22.55% in comparison to the Caiman either being Type I or Type II. Then we compute the test statistic for the sample percentage with the z-score critical value for the amount of 14.34% that will be added or subtracted from the sample percentage for our interval. This ends up giving us the interval (8.21, 36.89) for the proportion that we can be 95% confident the amount of Type II Caiman will be out of all Caiman either Type I or Type II.

Lastly, we are going to analyze the association between gender and Caiman Obesity. Here is the initial regression model with all explanatory variables. This is a worrying model. The p-value is high and the multiple R^2 is low.

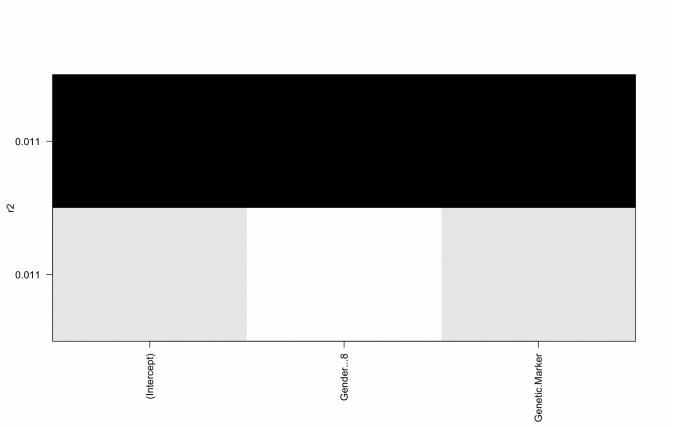
```
Residuals:
    Min      1Q  Median      3Q     Max
-2.6365 -1.7027  0.2289  1.1648  2.8257

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     7.19673    0.22961  31.344   <2e-16 ***
Gender...8      0.04156    0.31588   0.132    0.896
Genetic.Marker  0.38876    0.37677   1.032    0.305
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.59 on 99 degrees of freedom
Multiple R-squared:  0.01087,   Adjusted R-squared:  -0.009117
F-statistic: 0.5437 on 2 and 99 DF,  p-value: 0.5823
```

Next, we check autocorrelation, and we don't decide to remove any variables.

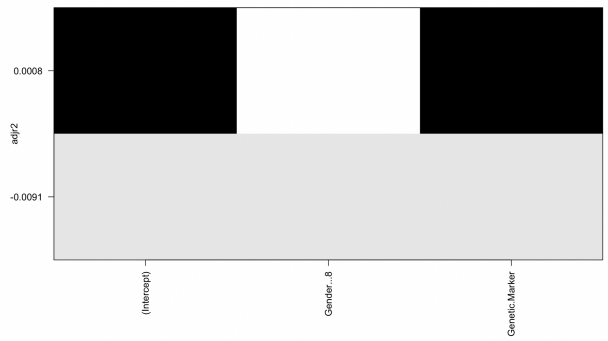
```
                Gender...8 Genetic.Marker
Gender...8      1.00000000     0.01891804
Genetic.Marker 0.01891804     1.00000000
```

Then, we decide to perform the subset selection method.

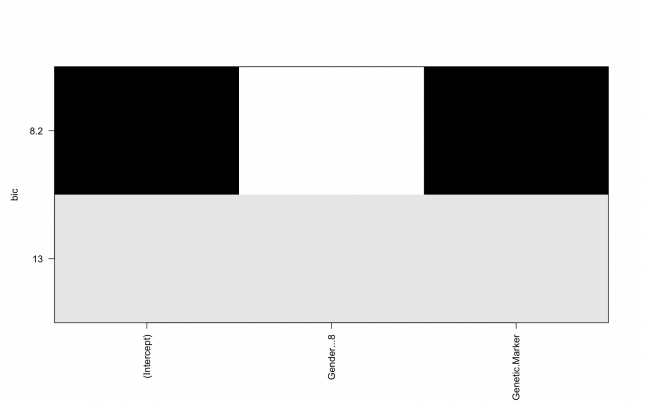
R^2:



Adjusted R^2:



BIC:

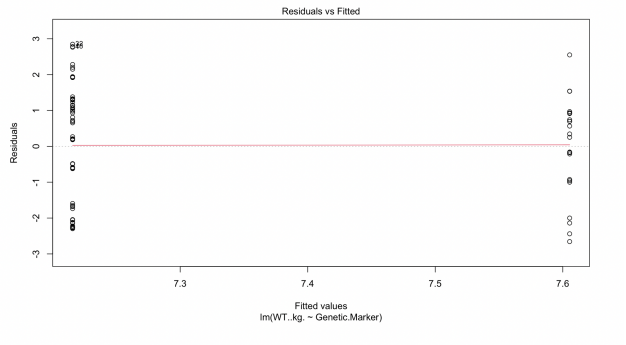


From the above graphs, due to the low probability and predictiveness of the Gender explanatory variable, we constitute the need for the removal of this variable. Now we compute a regression model with the explanatory variable of genetic markers (Type I and Type II) below. We do see the p-value has gone down for the model at 0.301, but it is still greater than we would hope. Also, the multiple R^2 value is 0.01069, which is slightly less than the already low numbers. This is very alarming. Furthermore, the F-statistic of 1.081 is very low, as well, and has only slightly increased from the first regression model.

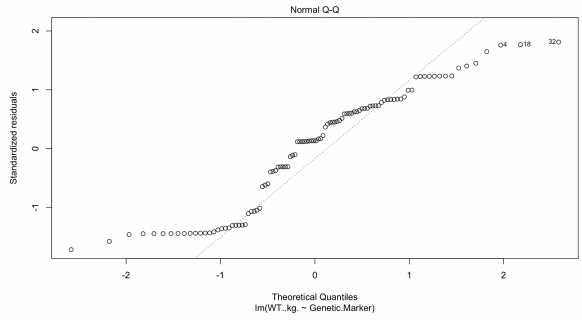
```
Residuals:
   Min     1Q Median     3Q    Max
-2.656 -1.680  0.210  1.146  2.848

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     7.2157     0.1780   40.54   <2e-16 ***
Genetic.Marker  0.3897     0.3748    1.04    0.301
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.582 on 100 degrees of freedom
Multiple R-squared:  0.01069,   Adjusted R-squared:  0.0007991
F-statistic: 1.081 on 1 and 100 DF,  p-value: 0.301
```

Lastly, we will check the assumptions:

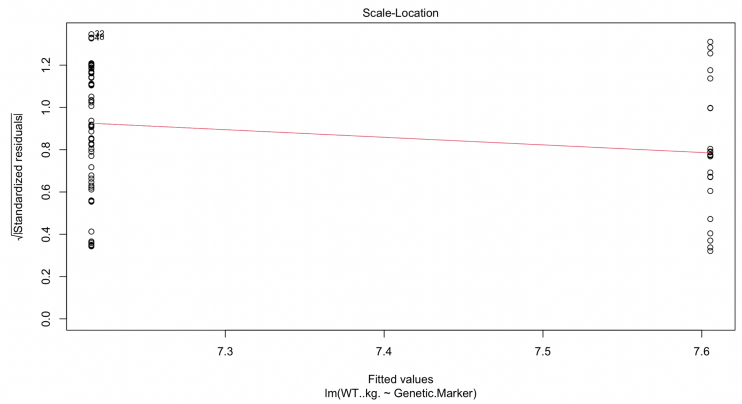
For the first assumption of linearity, we accept the assumption as the red line is horizontal. These are promising data values with a very similar residual difference.



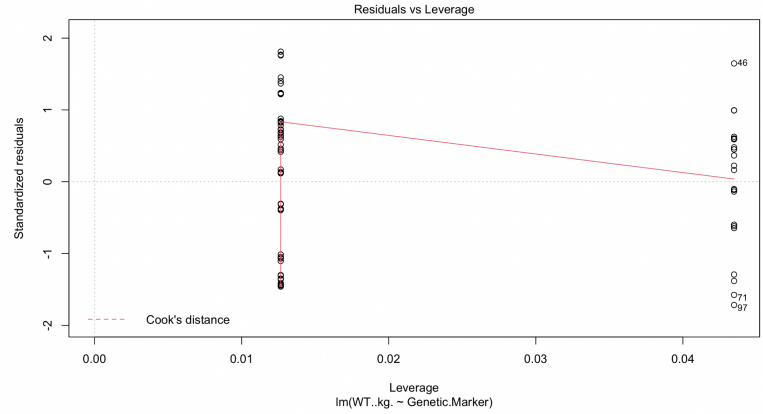
For the second assumption of normality, we reject the assumptions since the values don't appear to follow the line well. There appears to be movement of the data values away from the line.



For the third assumption of homoscedasticity, we accept the assumption because the red line is relatively horizontal. We see the standardized residuals are mostly similar.



For the fourth assumption of leverage, we accept the assumption since there appear to be no influential points. There are no data values past Cook's distance.



Overall, after noting that three of the four assumptions were accepted, that is promising. However, the predictability of the model is very low, and it isn't statistically significant at a p-value of 0.05. We also had to remove gender variables due to their ineffectiveness. This gives us the idea that obesity is not associated with gender. Furthermore, since we had to remove many variables and still produced little effect, we have a great distrust in the predictability of this

model with genetic markers (Type I or Type II) predicting Caiman weight in kg or any model of such coming from this data. Therefore, we highly disagree with the claims that Caiman obesity is the leading cause of tumor growth. If we were to change the model to weight as the explanatory and genetic markers as the response, it, unfortunately, produces similar results as shown below. The p-value, multiple R^2, and F-statistic are all the same in the above model.

```
    Min      1Q  Median      3Q     Max
-0.3012 -0.2538 -0.2082 -0.1608  0.8391

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02510    0.19719   0.127    0.899
WT..kg.      0.02744    0.02639   1.040    0.301

Residual standard error: 0.4198 on 100 degrees of freedom
Multiple R-squared:  0.01069,   Adjusted R-squared:  0.0007991
F-statistic: 1.081 on 1 and 100 DF,  p-value: 0.301
```

4. In this section, we are analyzing the CDC gut bacteria of average Saccharomyces diameter (micro m) and Lactobacillus-X ppm and its potential explanatory variables of regions (Africa, North and South America, Asia, and Europe), weight, height, and cholesterol. For the regions, we will have explanatory variables split up into dummy variables of D1 (relating 0 as Africa and 1 as North and South America), D2 (relating 0 as not Asia and 1 as Asia), D3 (relating 0 as not Europe and 1 as Europe).

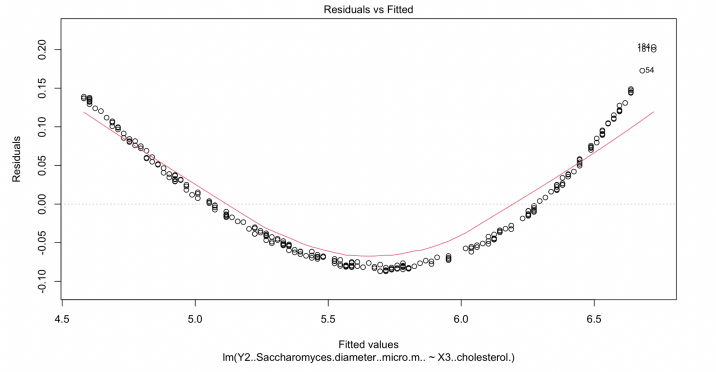
For our first task, we are analyzing if there is a relationship that exists between cholesterol levels and average Saccharomyces diameter. First, we will implement the regression model.

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.08700 -0.06701 -0.01646  0.05530  0.20333

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      10.3652778  0.0359485   288.3   <2e-16 ***
X3..cholesterol. -0.0214224  0.0001611  -133.0   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07357 on 250 degrees of freedom
Multiple R-squared:  0.9861,    Adjusted R-squared:  0.986
F-statistic: 1.768e+04 on 1 and 250 DF,  p-value: < 2.2e-16
```

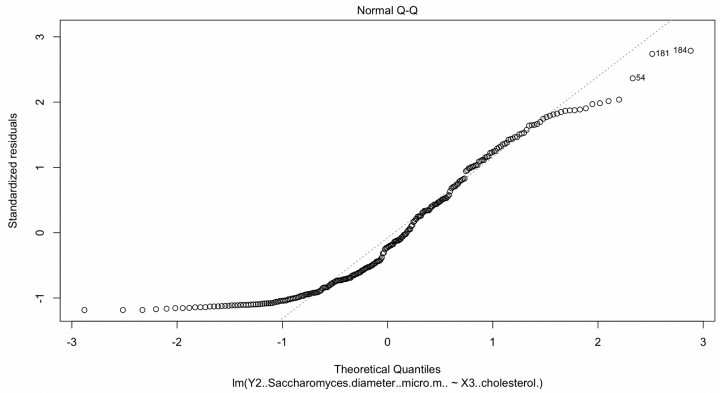
From this regression model, we see that the p-value is very low and statistically significant for a p-value of less than 0.05. The predictability of the model appears to be good as well with a multiple R^2 value of 0.9861. Lastly, the F-statistic appears to be very high as well. Overall, this appears to be a promising relationship in the regression model.

Next, we will skip checking for autocorrelation and performing the subset selection method since we only have one variable to work with. We will proceed with checking our assumptions.

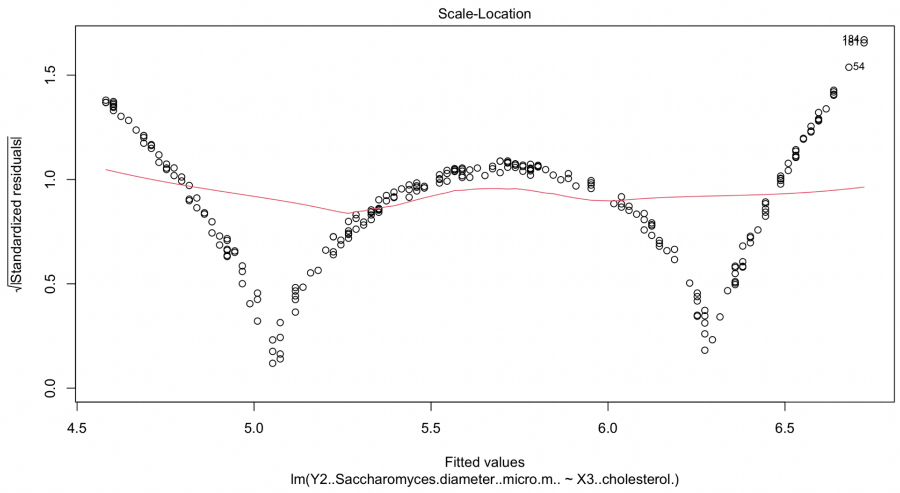
First, we check the assumption of linearity. This assumption appears to not be followed due to the red line not being horizontal. The data values make a u-shape that shows no similar difference in residuals on either side of the line.



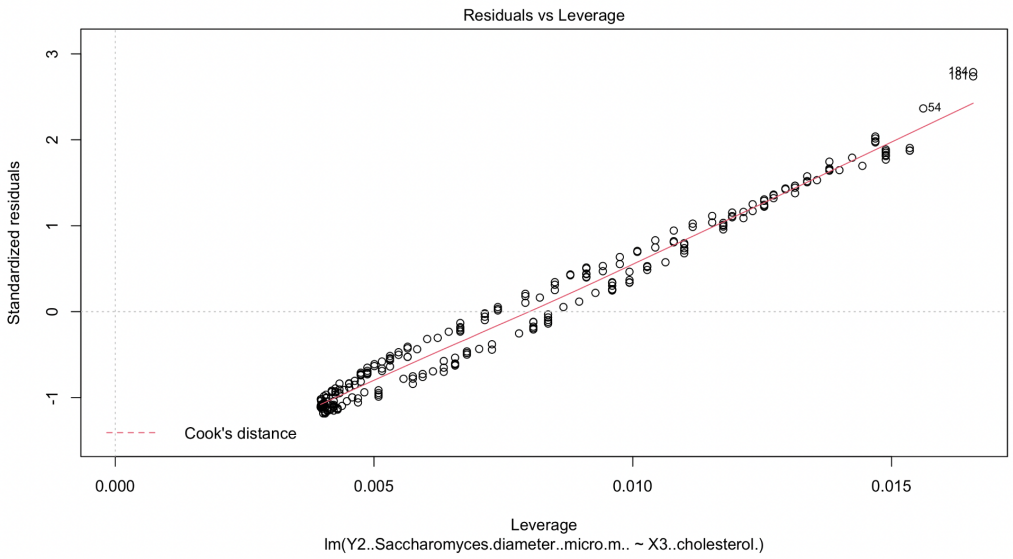
For the second assumption of normality, we relatively accept the assumption since the values mostly follow the line. There is some movement of values away from the line, but it isn't enough to reject the assumption.



For the third assumption of homoscedasticity, we accept the assumption since the red line appears to be horizontal. We can see the standardized residuals tend to follow in a movement of values that is symmetric throughout.



For the fourth assumption of leverage, we accept the assumption due to there being no influential points. There are no values past Cook's distance.



Overall, three out of four assumptions appeared to be followed, and the model displays great predictability. Therefore, we can conclude that there is a possible relationship between cholesterol and average Saccharomyces diameter.

For the next topic, we are looking at the typical regression model used to predict Lactobacillus-X. Below is the result of this regression model. We see that the p-value is statistically significant by being lower than 0.05. Also, the multiple R^2 values (0.9475) and the F-statistic (736.8) are very high. This model appears to be promising, but we need to look at the autocorrelation of the variables.

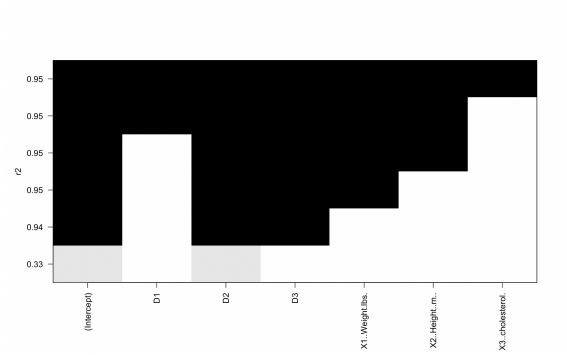
```
Residuals:
    Min      1Q  Median      3Q     Max
-63.790 -12.089  -1.323  10.653  61.523

Coefficients:
                  Estimate Std. Error t value   Pr(>|t|)
(Intercept)       264.33902   20.39619  12.960    < 2e-16 ***
D1                  1.88318    5.16183   0.365     0.7156
D2               -222.69587    4.49135 -49.583    < 2e-16 ***
D3               -223.97863    4.73276 -47.325    < 2e-16 ***
X1..Weight.lbs.     0.26940    0.05398   4.991 0.00000114 ***
X2..Height..m..    17.21783    8.75077   1.968     0.0502 .
X3..cholesterol.   -0.00712    0.05894  -0.121     0.9039
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.68 on 245 degrees of freedom
Multiple R-squared:  0.9475,    Adjusted R-squared:  0.9462
F-statistic: 736.8 on 6 and 245 DF,  p-value: < 2.2e-16
```

Autocorrelation:

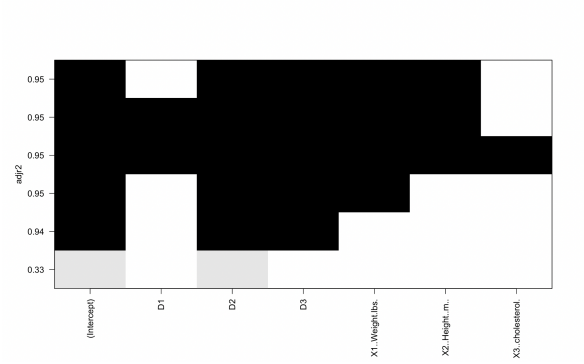
```
                          D1          D2          D3 X1..Weight.lbs.
D1                 1.00000000 -0.31638653 -0.27282532     -0.01214226
D2                -0.31638653  1.00000000 -0.38655567     -0.01790263
D3                -0.27282532 -0.38655567  1.00000000      0.01667341
X1..Weight.lbs.   -0.01214226 -0.01790263  0.01667341      1.00000000
X2..Height..m..   -0.02773102 -0.01955679 -0.03130527     -0.06686218
X3..cholesterol. -0.09740767  0.02486529  0.06323478     -0.01221659
                  X2..Height..m.. X3..cholesterol.
D1                     -0.02773102      -0.09740767
D2                     -0.01955679       0.02486529
D3                     -0.03130527       0.06323478
X1..Weight.lbs.        -0.06686218      -0.01221659
X2..Height..m..         1.00000000       0.07665295
X3..cholesterol.        0.07665295       1.00000000
```

From this, we see that there don't appear to be any relationships between the variables that constitute removal. We will proceed with the subset selection method with the remaining variables.

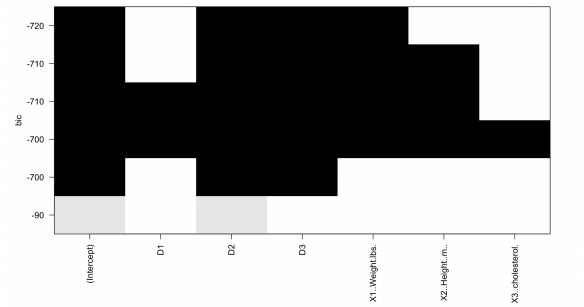
R^2:



Adjusted R^2:



BIC:



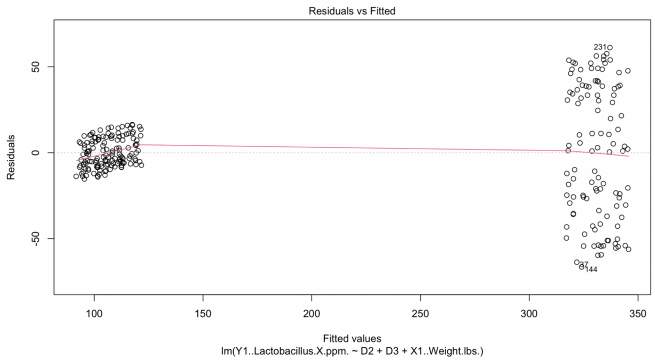
From the graphs above, we constitute the removal of D1(of Africa 0 and North and South America 1), Height, and Cholesterol as explanatory variables. We will create a regression model below with the remaining variables of Weight, D2(with not Asia 0 or Asia 1), and D3(with not

Europe 0 or Europe 1). This model has the same p-value, a slightly reduced multiple R^2 to 0.9466, and an increased F-statistic to 1467.

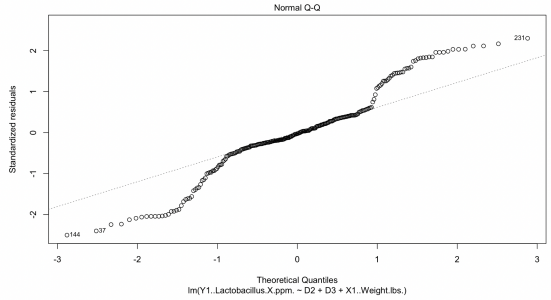
```
Residuals:
    Min      1Q  Median      3Q     Max
-66.466 -10.541  -0.696  11.147  61.155

Coefficients:
                 Estimate Std. Error t value   Pr(>|t|)
(Intercept)      290.46840    8.69344  33.412    < 2e-16 ***
D2              -223.77714    3.95019 -56.650    < 2e-16 ***
D3              -225.14535    4.21725 -53.387    < 2e-16 ***
X1..Weight.lbs.    0.26204    0.05395   4.857 0.00000211 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.73 on 248 degrees of freedom
Multiple R-squared:  0.9466,    Adjusted R-squared:  0.946
F-statistic:  1467 on 3 and 248 DF,  p-value: < 2.2e-16
```

Lastly, we will check the assumptions for the model.

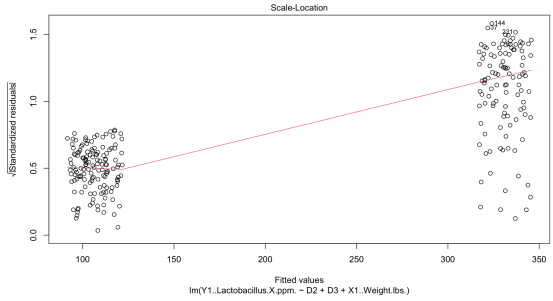
For the first assumption of linearity, we accept the assumption due to the red line being horizontal. We see the values are slightly different but maintain the horizontal line.



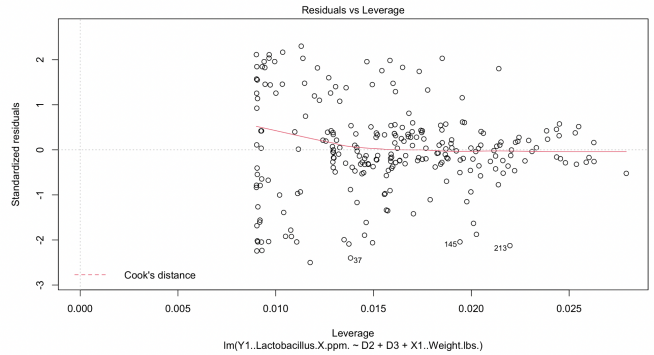
For the second assumption of normality, we reject the assumption due to the values not consistently following the line and oscillating. We see this oscillation at the beginning and end of the standardized residual amounts.



For the third assumption of homoscedasticity, we reject the assumption since the line is not directly horizontal. The end values of the standardized residuals are higher than the beginning.

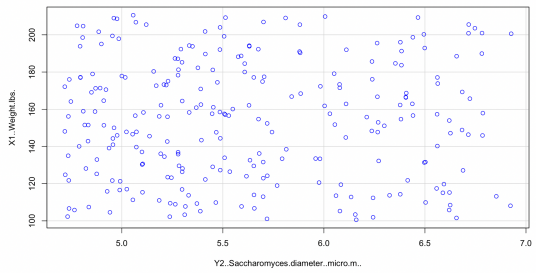


For the fourth assumption of leverage, we accept the assumption without the appearance of influential points. There are no values past Cook's distance.



We see that two of the four assumptions were followed in this regression model. Overall, the initial regression model is okay on its own; however, it could reduce some of its variables to increase its effectiveness. We believe the important explanatory variables to be D2(Asia), D3(Europe), and Weight in lbs for predicting Lactobacillus-X ppm specifically.

For our last analysis, we will assess if large-diameter Saccharomyces are associated with low weight. We initially create a scatterplot below to display the data, from this, we don't see there to be any association between the two variables. However, we will create a regression model to assess the possibility.

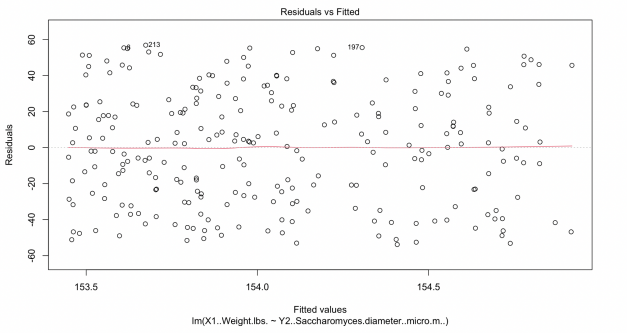


Here is the regression model below. From this, we can see that the p-value is very high at 0.8345. Also, the predictability of the model is very low with a multiple R^2 value of 0.000175. Lastly, the F-statistic is very low at 0.04377. Therefore, the model appears to be showing there is no relationship.

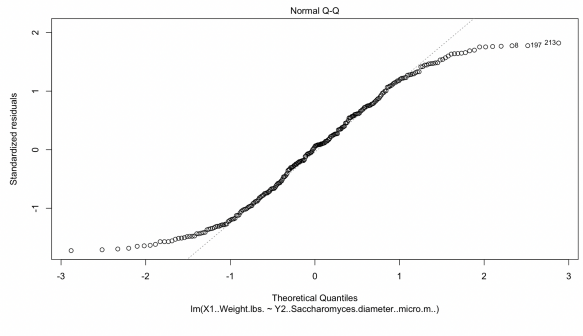
```
Residuals:
    Min      1Q  Median      3Q     Max
-53.846 -26.882   1.676  24.285  56.875

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                         150.3083    18.0052   8.348 4.77e-15 ***
Y2..Saccharomyces.diameter..micro.m..   0.6656     3.1816   0.209    0.834
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31.35 on 250 degrees of freedom
Multiple R-squared:  0.000175,  Adjusted R-squared:  -0.003824
F-statistic: 0.04377 on 1 and 250 DF,  p-value: 0.8345
```

However, we will still check our assumptions.

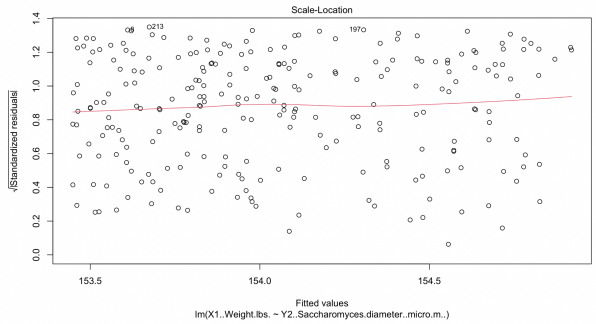
For the first assumption of linearity, it is accepted with a horizontal red line. We see a similar occurrence in residuals on either side of the line.



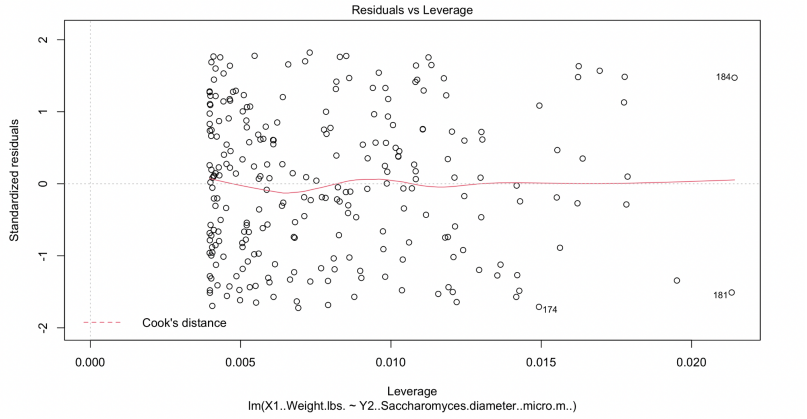
The second assumption of normality appears to be mostly following the line. So we decide to accept this assumption. It is only slightly at the beginning and end that the data values move away from the line.



The third assumption of homoscedasticity appears to be followed with a relatively horizontal red line. We see the data values are relatively similar on either side of the line.



For the fourth assumption of leverage, we accept the assumption with no influential points. There are no values that go past Cook's distance.



Overall, we accepted all four assumptions. However, the model predictability is still worrying, and the relative scatter plot shows a very minimal relationship between the variables. Therefore, we can't conclude that large-diameter Saccharomyces in micro m is associated with low weight in lbs. Even further, we have reason to believe there isn't a relationship between the two variables. Use this regression model with caution.