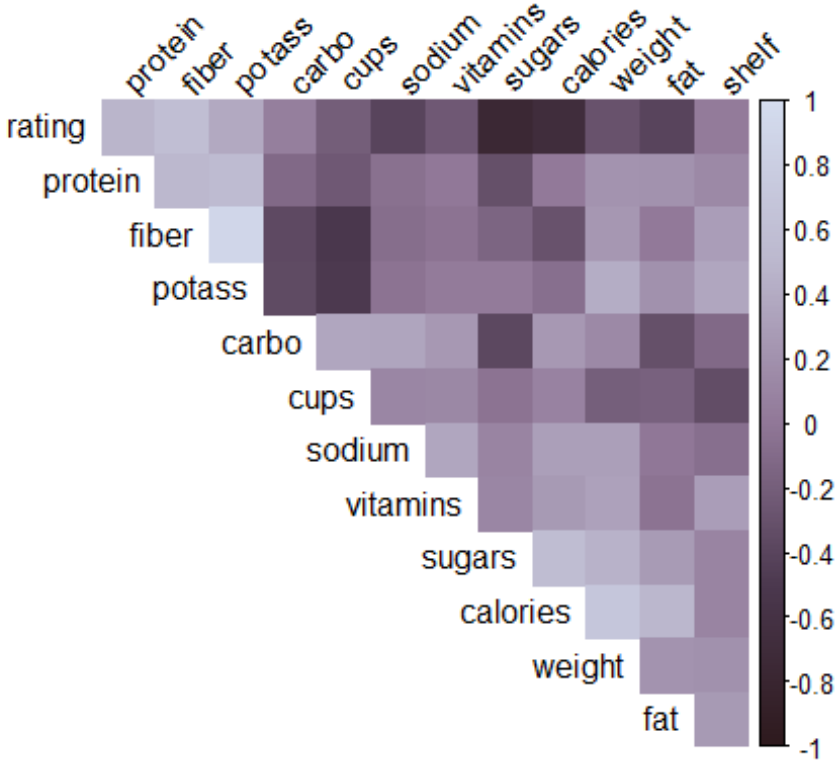
Cereals

Made by majafoi

UPDATED 7/5/2021
MADE IN R/R studio

```
cereals = read.csv ("cereal-cor.csv") # to load our dataset called Cereals
attach (cereals) # to omit the need of writing dataset-name$variable-name

library (corrplot)

## corrplot 0.84 loaded

library (ggpubr)

## Loading required package: ggplot2

cereals.correlation = cor (cereals)
col = colorRampPalette (c("#2e1a1e","#4c394f","#917898","#bcb8ce","#d5ddef"))
corrplot (cereals.correlation, method = "color", type = "upper", order = "hclust", addCoef.col = NULL, col = col(200), tl.col = "
black", tl.srt = 45, sig.level = 0.01, insig = "blank", diag = FALSE)

## Warning in ind1:ind2: numerical expression has 2 elements: only the first used
```

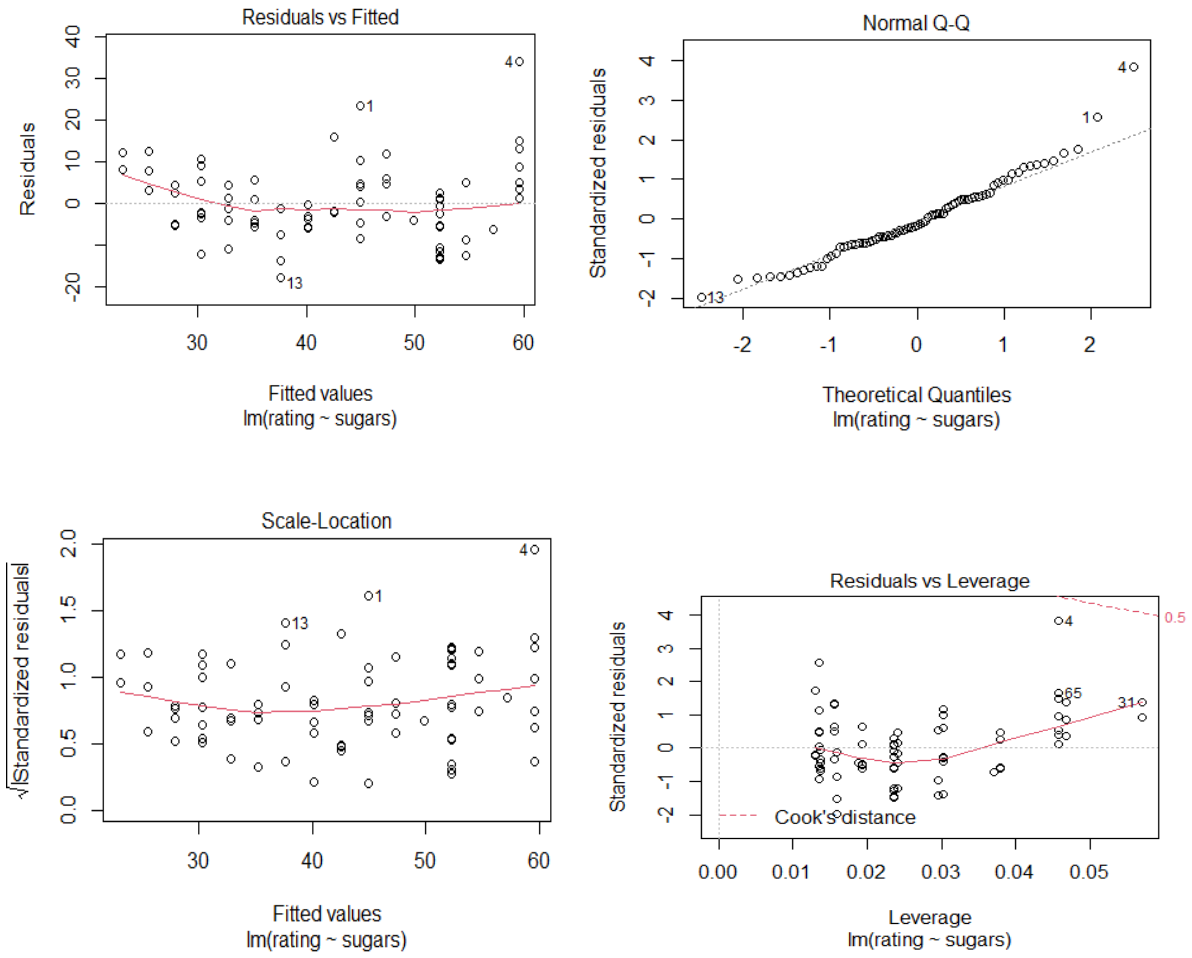
Here you can see the correlation matrix of the variables in the sample. The darker the color, the stronger is the correlation between those two variables. Since we are interested in predicting the rating, I will only look at the first row. It appears that rating is highly correlated with the variables sodium, sugars, calories and fat, so I'll only check and use those variables. Since I want to predict the variable, I'll use regression analysis.



```
reg1 = lm (rating ~ sugars) # this is the function for regression model rating = a*sugars + b
print (summary (reg1))

##
## Call:
## lm(formula = rating ~ sugars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.849  -5.619  -1.359   4.926  34.117
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  59.584     1.954   30.50  < 2e-16 ***
## sugars       -2.435     0.238  -10.23 6.99e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.135 on 75 degrees of freedom
## Multiple R-squared:  0.5827, Adjusted R-squared:  0.5771
## F-statistic: 104.7 on 1 and 75 DF,  p-value: 6.992e-16

plot(reg1)
```

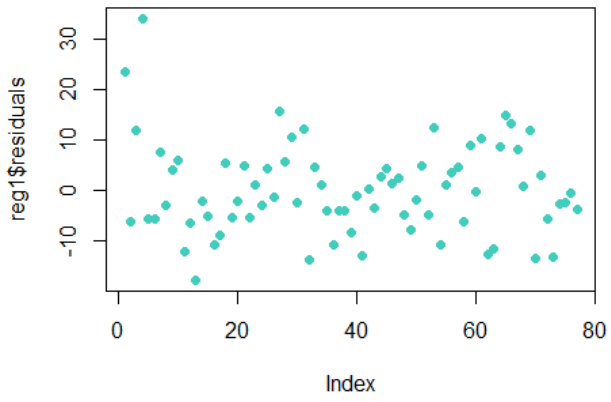
R2 is coefficient of determination or the proportion of the total variability which is explained by the model itself (the variables in it). If you are trying to predict something, then you are bound to make some mistakes. Those "mistakes" or leftovers here are called residuals (the difference between the real model and predicted one). The key is to "grab" as much of those residuals in the model, aka to get as high as possible R2. Everything above 10% of the variability in the model is okay, although the number I would be happy with is over 50% with one variable.

Our adjusted R2 is 57.71%, which is a nice percentage to continue with. Next part of the model is p-value. The main hypothesis is that the variable we are looking at is not important for the model. For us to refuse that hypothesis, we need the p-value of the variable and the model to be lower than 0.05 (we are 95% certain we are right about it). Here, both p-values are lower than 0.05, and we can say that the variable sugars is important for the prediction model (for variable rating). The only potential problem here are the residuals. Above in the result, you can see the distribution of the residuals, and those numbers should be close to zero. Since they aren't, I'm going to check the plot of the residuals. In this case, they shouldn't have any particular shape, but they should be scattered all over the graph.



Long story short, the q-q diagram should show a pretty straight line which would show that both sets of quantiles (both rating and sugars) come from normal distributions. Our graph looks like it, with 3 observations that are clearly outliers. But, we should also look at the Residuals vs Leverage plot (bottom right). It shows us if any point have substantial influence on the regression model (if so, those should be outside the Cook's distance of 0.5. None is here, although one is pretty close. The plot below shows us if the residuals are scattered. To me, those are scattered residuals, and we can continue with the regression model.

```
plot (reg1$residuals, pch=16, col="#41CCBC")
```

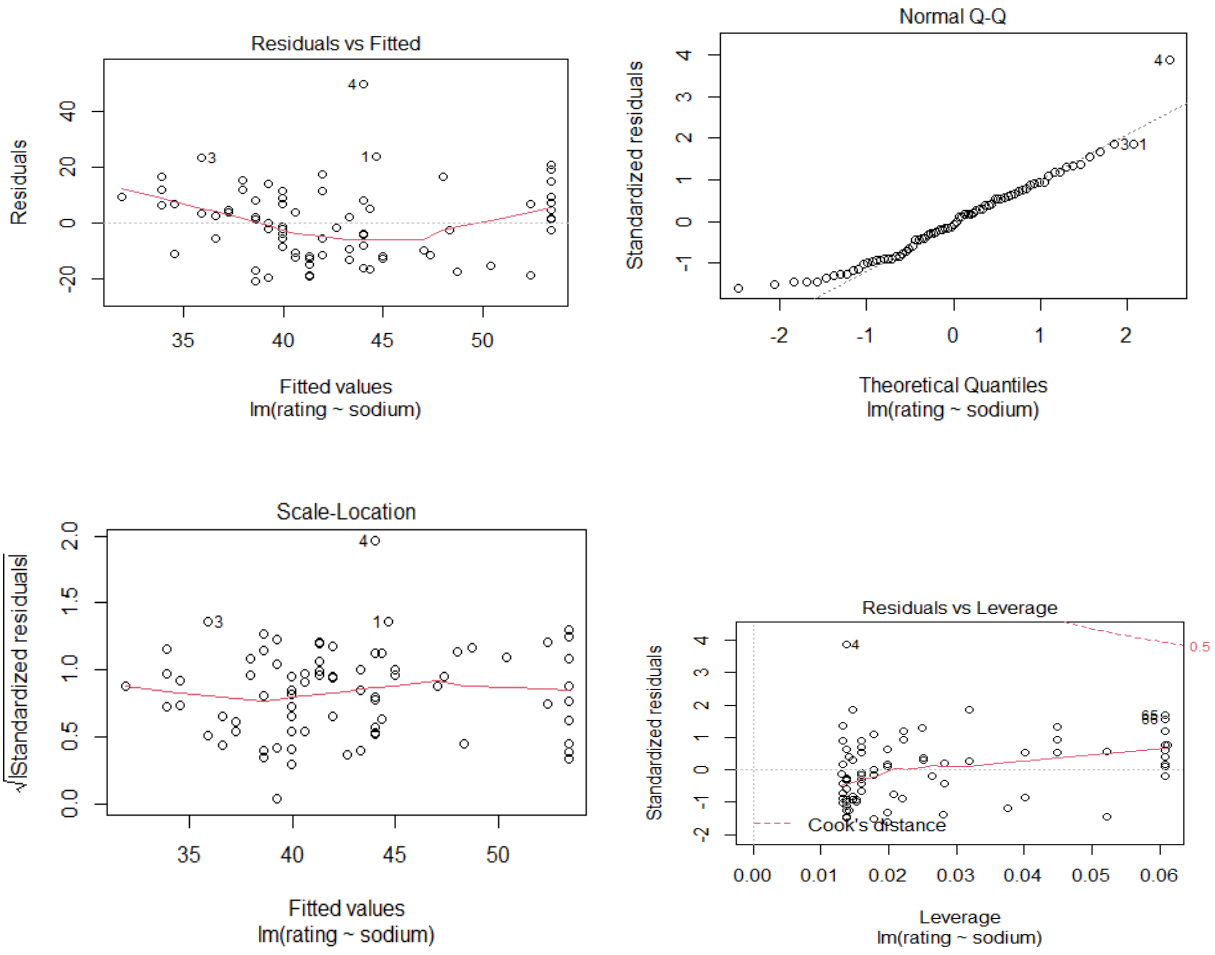


The regression model is y=ax+b OR rating = 59,584 – 2,435 * sugars

Let's say that the sugar level in the cereal is 6,948g (that is the average value). The rating then is → rating = 59,584 – 2,435 * 6,948 = 42.66.

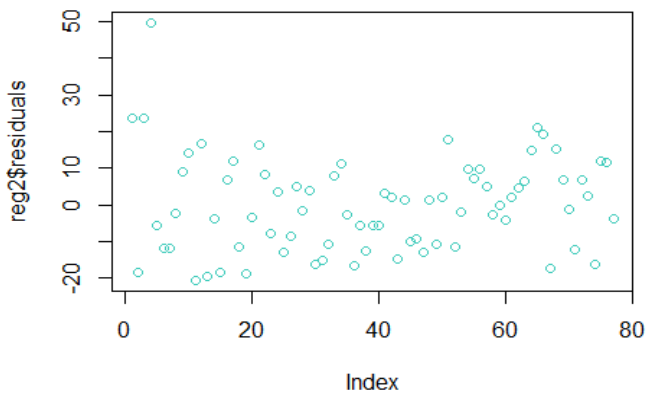
Next regression model we have is rating and sodium level. Both p-values are significant, but the R2 is quite low – only 14.99%. Since the distribution of residuals isn't close to zero, let's check the graphs.

```
reg2 = lm (rating~sodium)
print (summary(reg2))

##
## Call:
## lm(formula = rating ~ sodium)
##
## Residuals:
##    Min     1Q  Median     3Q    Max
## -20.569 -10.774  -1.114   8.092  49.712
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 53.40251    3.19163  16.732  < 2e-16 ***
## sodium      -0.06724    0.01772  -3.794 0.000298 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.95 on 75 degrees of freedom
## Multiple R-squared:  0.161,  Adjusted R-squared:  0.1499
## F-statistic:  14.4 on 1 and 75 DF,  p-value: 0.0002979

plot (reg2)
```



The q-q diagram looks normal like the previous one, with only one outlier, but the leverage graph looks a bit different. It doesn't have the "linear" look as it should have and it looks like that some other observation is having a big leverage (55).

```
plot(reg2$residuals,col="#41CCBC")
```



But again, the scatter plot of the residuals looks random, so let's proceed with the regression model.

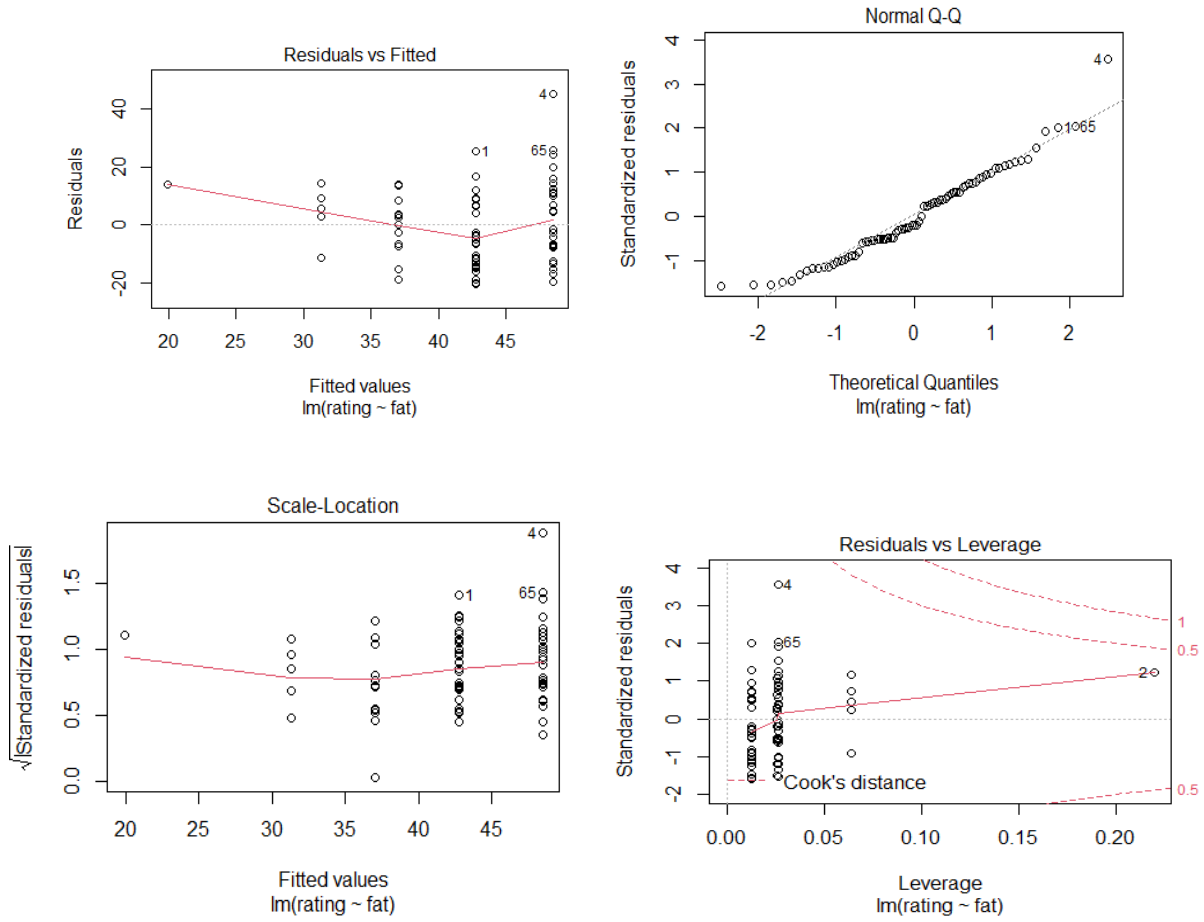
The regression model is y=ax+b OR **rating = 53.4025 – 0,06724 * sodium**

Let's say that the sugar level in the cereal is 159,7mg (that is the average value). The rating then is → **rating = 53.4025 – 0,06724 * 159,7 = 42,66**

Interesting. The rating is the same as in the previous regression model.

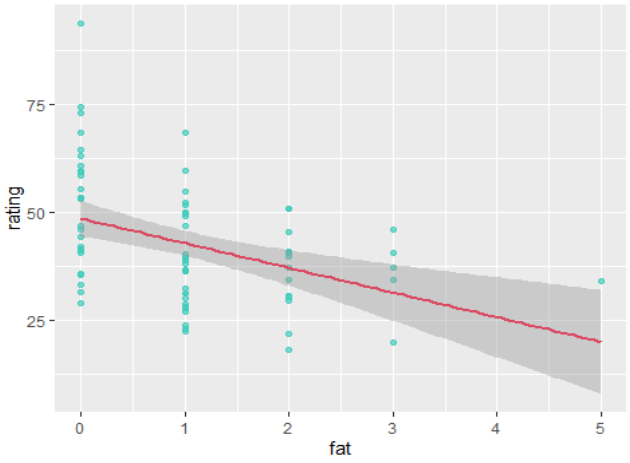
The next model is between rating and fat level. P-values are significant, the R2 is low again – 15.64%.

```
reg3 = lm (rating~fat)
print (summary(reg3))

##
## Call:
## lm(formula = rating ~ fat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.340  -7.892  -2.630   8.850  45.248
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   48.452      2.093  23.150  < 2e-16 ***
## fat           -5.713      1.470  -3.885 0.000219 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.9 on 75 degrees of freedom
## Multiple R-squared:  0.1675, Adjusted R-squared:  0.1564
## F-statistic: 15.09 on 1 and 75 DF,  p-value: 0.0002189

plot (reg3)
```

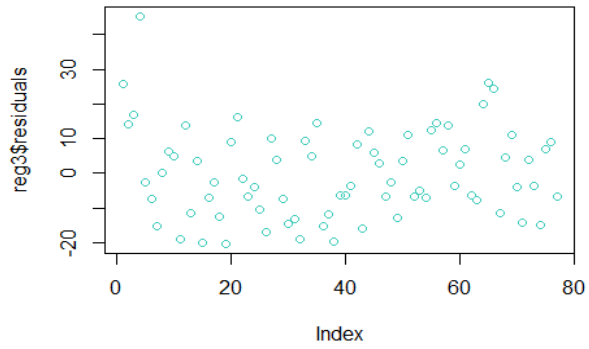


The q-q- diagram looks like its fine, but the leverage diagram isn't, because one observation really does have a huge leverage. Also, rating and fat variables don't have linear relationship. I won't be continuing with this model, although the residuals are randomly scattered.

```
ggplot(cereals, aes(x = fat, y = rating)) + geom_point(alpha=.7, color = "#41CCBC") + stat_smooth(method = lm, level = 0.95, color = "#DA4A62")

## `geom_smooth()` using formula 'y ~ x'
```

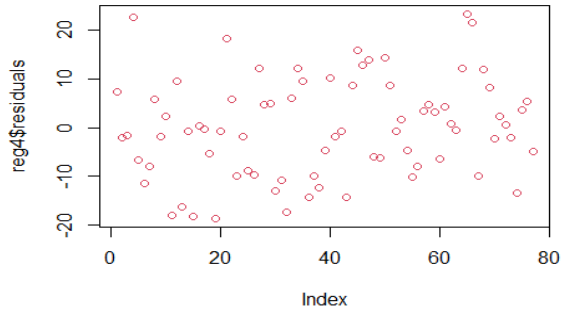



```
plot(reg3$residuals, col = "#41CCBC")
```



The last individual regression model is between rating and calories. Both p-values are significant, with R2 being 46.83% which is a decent percentage.

```
reg4=lm(rating~calories)
print(summary(reg4))

##
## Call:
## lm(formula = rating ~ calories)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.7161  -7.9278  -0.6661   5.9936  23.4133
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  95.7892     6.5504  14.623  < 2e-16 ***
## calories     -0.4970     0.0603  -8.242 4.13e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.24 on 75 degrees of freedom
## Multiple R-squared:  0.4753, Adjusted R-squared:  0.4683
## F-statistic: 67.93 on 1 and 75 DF,  p-value: 4.132e-12

plot(reg4$residuals,col="#DA4A62")
```



```
plot (reg4)
```

Q-Q diagram is normal, but the leverage diagram shows that some values have a big degree of leverage. The residuals are scattered randomly, but I won't be continuing with this model.