R Notebook

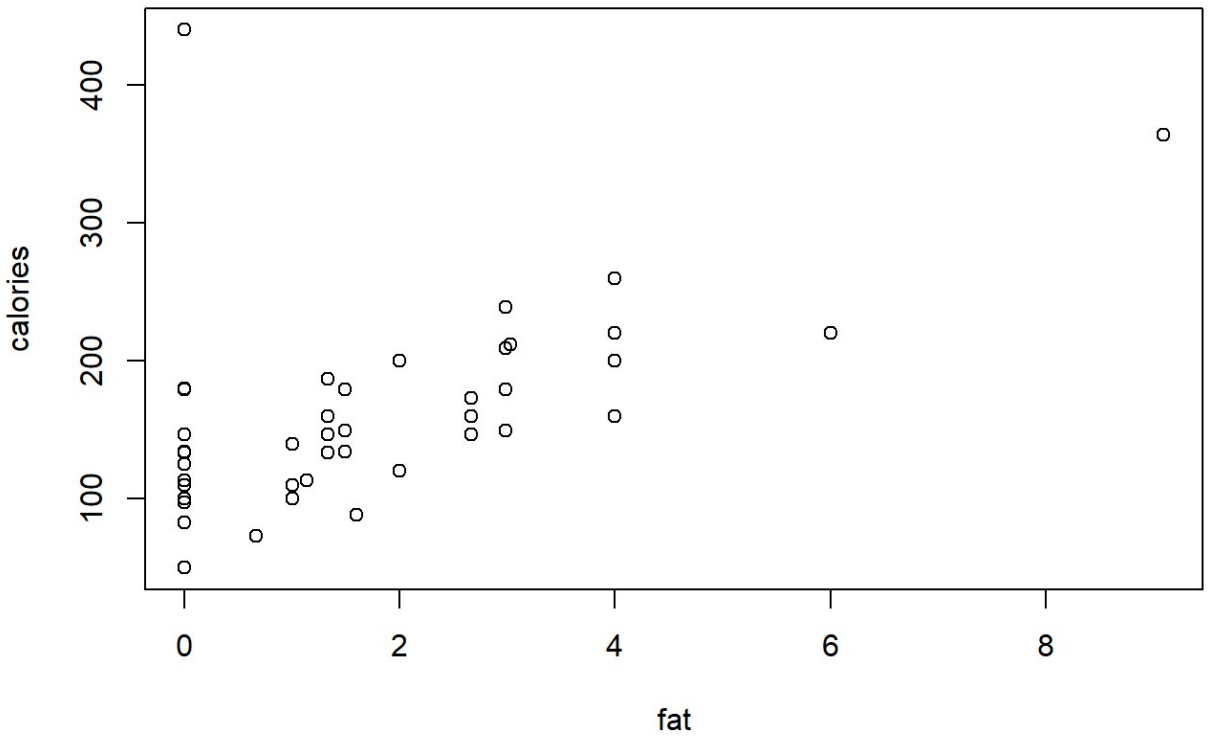
```
rm(list=ls(all=TRUE))

library(MASS)
data(UScereal)
UScereal
```

	mfr <fctr>	calories <dbl>	protein <dbl>	fat <dbl>
100% Bran	N	212.12121	12.1212121	3.0303030
All-Bran	K	212.12121	12.1212121	3.0303030
All-Bran with Extra Fiber	K	100.00000	8.0000000	0.0000000
Apple Cinnamon Cheerios	G	146.66667	2.6666667	2.6666667
Apple Jacks	K	110.00000	2.0000000	0.0000000
Basic 4	G	173.33333	4.0000000	2.6666667
Bran Chex	R	134.32836	2.9850746	1.4925373
Bran Flakes	P	134.32836	4.4776119	0.0000000
Cap'n'Crunch	Q	160.00000	1.3333333	2.6666667
Cheerios	G	88.00000	4.8000000	1.6000000

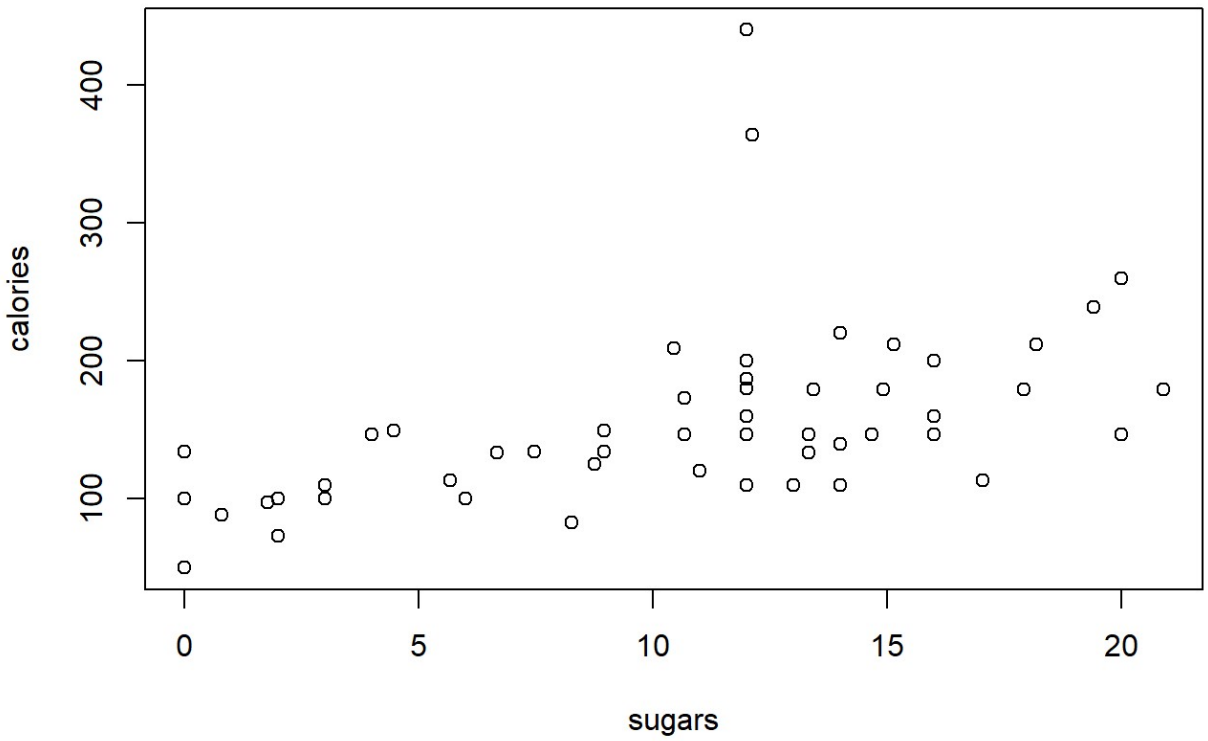
1-10 of 65 rows | 1-7 of 12 columns Previous **1** 2 3 4 5 6 7 Next

First, we clear the memory as well as import the data that we'll be using in this assignment. Next, we determine the alternative hypothesis as well as the null hypothesis. In this scenario, we pose the research question: "Does the quantity of molecules such as fats, sugars, complex carbohydrates, proteins, etc. determine the number of calories in a cereal?" The null hypothesis is as follows: "The quantity of molecules does not determine the number of calories in a cereal". The alternative hypothesis is as follows: "The quantity of molecules determines the number of caloies in a cereal".

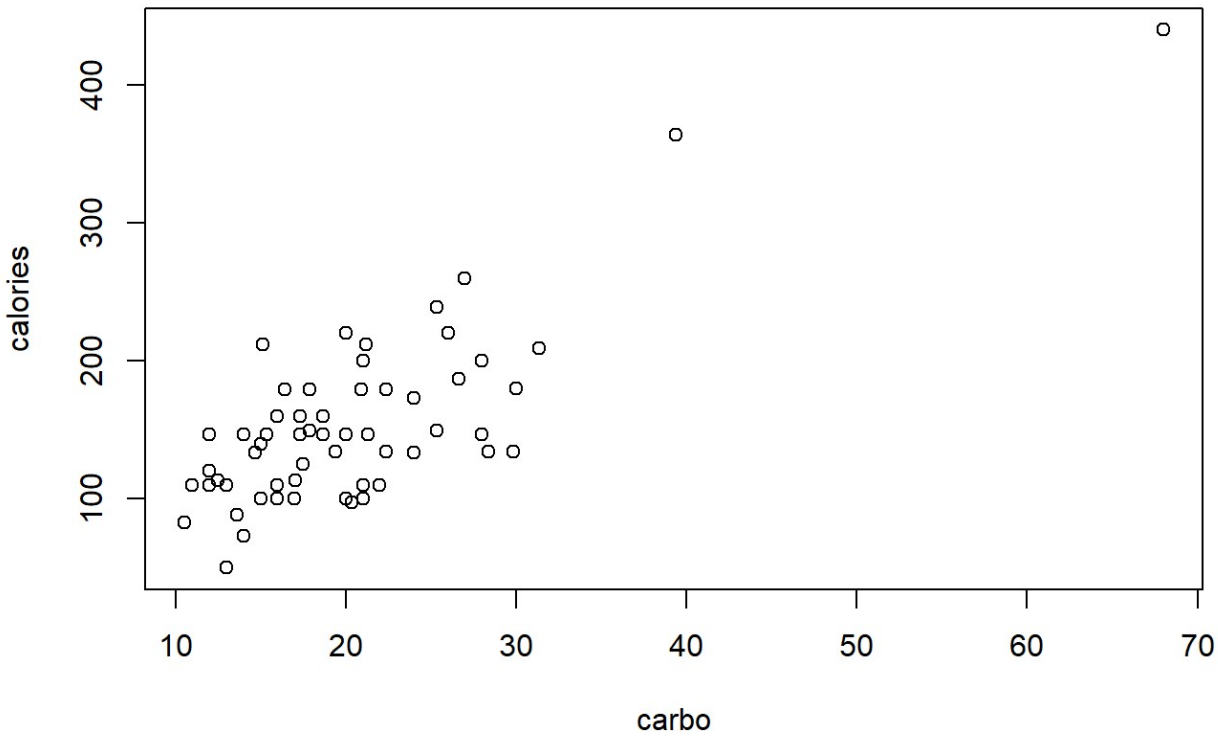
```
plot(calories ~ fat, UScereal)
```



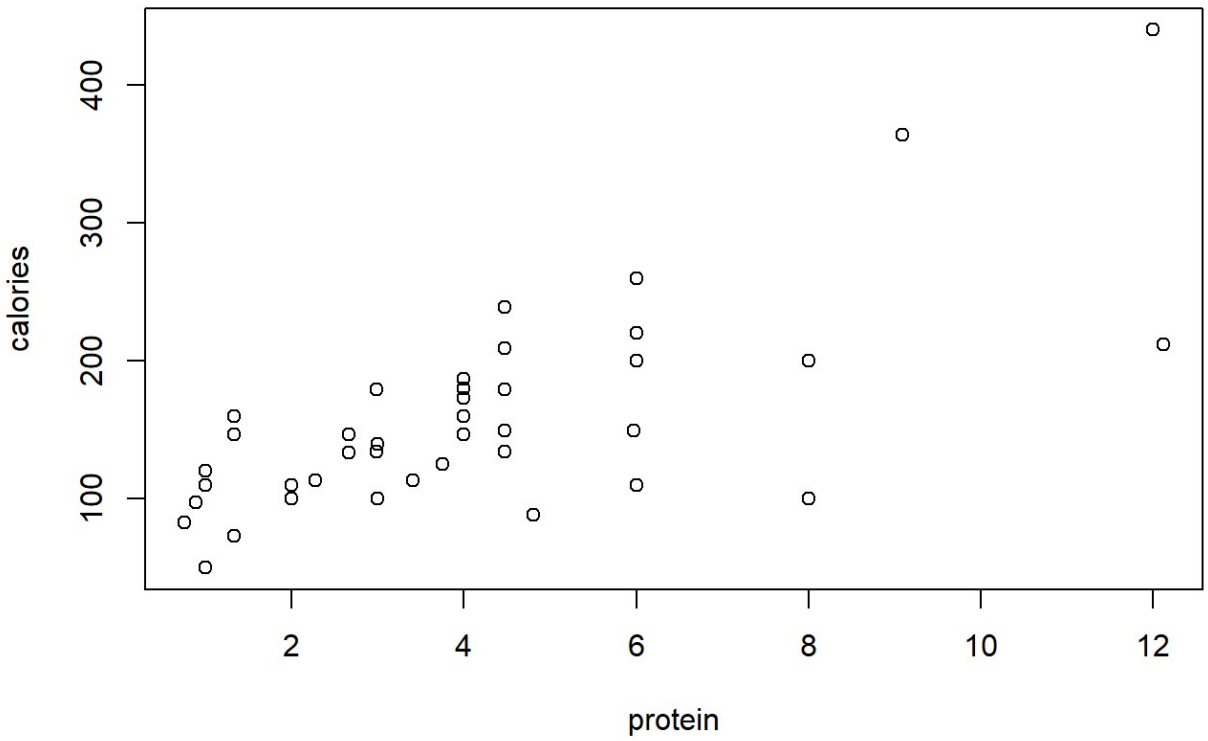

```
plot(calories ~ sugars, UScereal)
```




```
plot(calories ~ carbo, UScereal)
```




```
plot(calories ~ protein, UScereal)
```

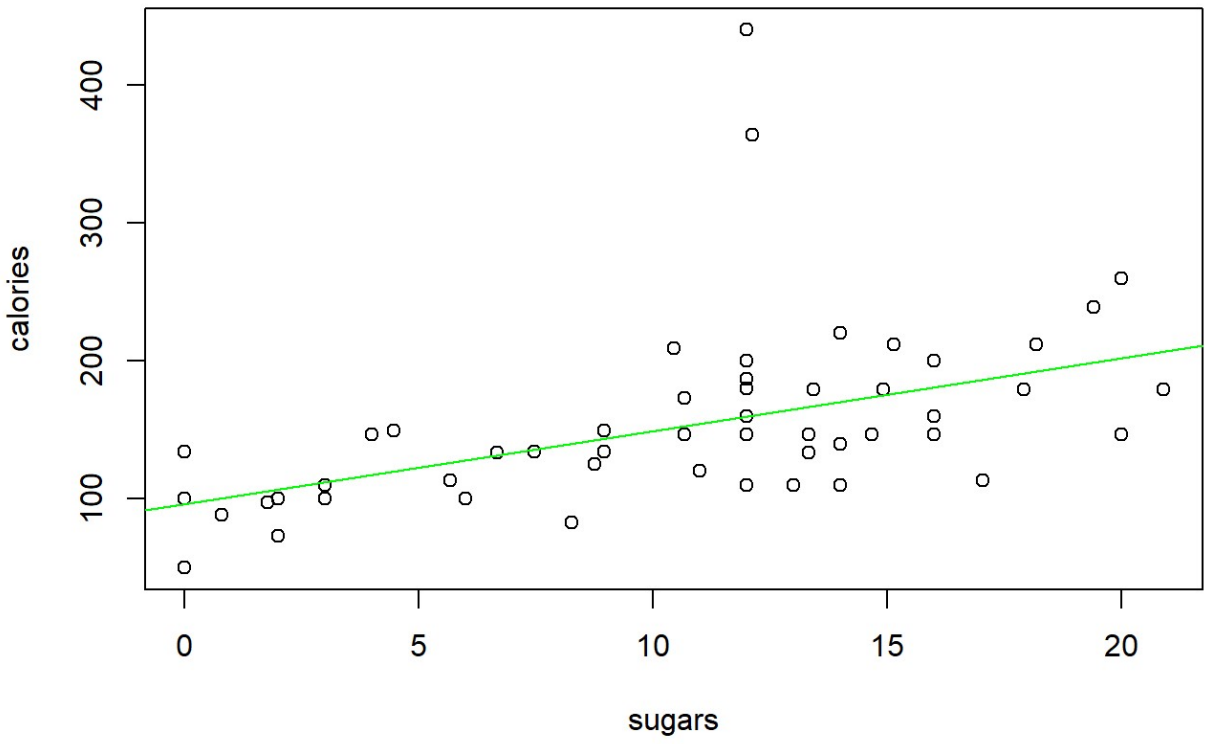



```
cereal <- lm(calories ~ fat + sugars + carbo + protein + fat:sugars + fat:carbo + fat:
protein + sugars:carbo + sugars:protein + carbo:protein, data = UScereal)

summary(cereal)
```

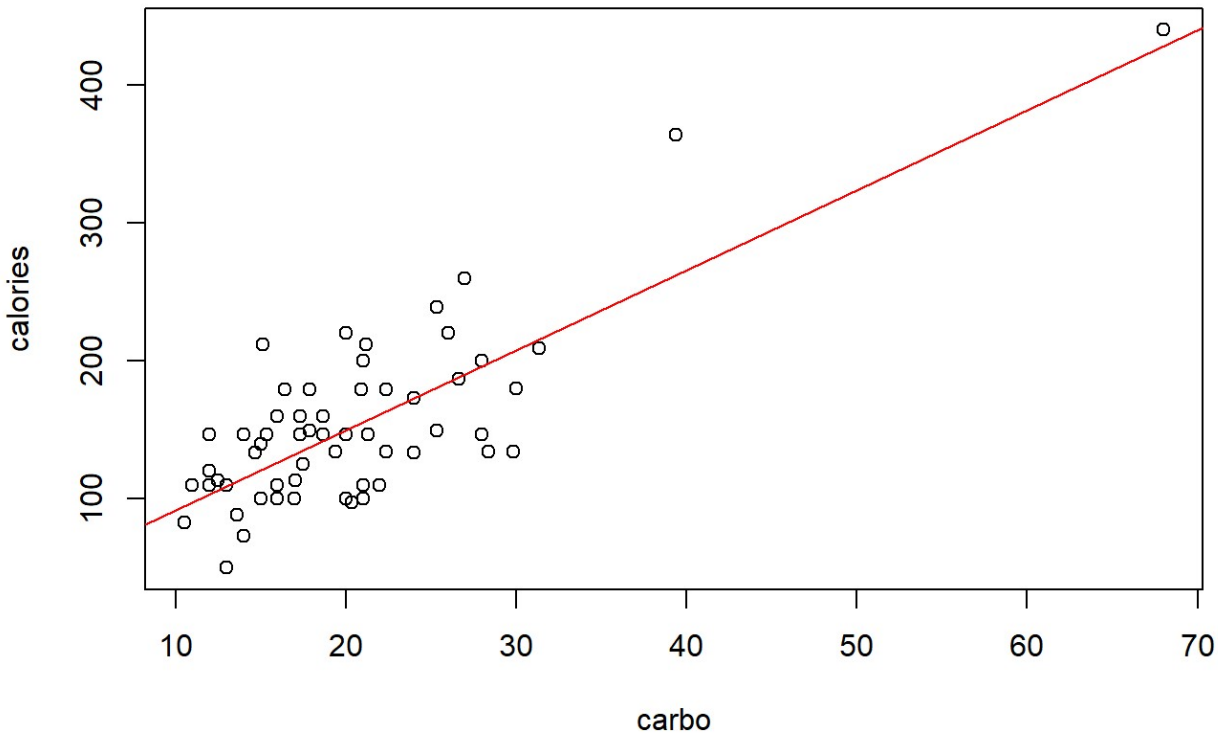
```
##
## Call:
## lm(formula = calories ~ fat + sugars + carbo + protein + fat:sugars +
##     fat:carbo + fat:protein + sugars:carbo + sugars:protein +
##     carbo:protein, data = UScereal)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.2642  -3.7979   0.9519   4.5526  12.4044
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    18.58096    7.20031   2.581  0.01261 *
## fat             1.39064    2.48357   0.560  0.57784
## sugars          3.25487    0.55861   5.827 3.23e-07 ***
## carbo           3.08819    0.33207   9.300 8.37e-13 ***
## protein         1.42788    1.00551   1.420  0.16134
## fat:sugars      0.42144    0.15291   2.756  0.00796 **
## fat:carbo       0.09039    0.11381   0.794  0.43054
## fat:protein     0.07779    0.44342   0.175  0.86140
## sugars:carbo    0.03631    0.03082   1.178  0.24399
## sugars:protein -0.07505    0.09729  -0.771  0.44382
## carbo:protein   0.16311    0.02667   6.117 1.11e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.474 on 54 degrees of freedom
## Multiple R-squared:  0.9909, Adjusted R-squared:  0.9892
## F-statistic: 589.5 on 10 and 54 DF,  p-value: < 2.2e-16
```

Next, we take a look at the independent relationships as well as the casual pairwise relationships using the linear model, lm() function. The console then gives us the coefficients for the relationship as well as the p-values to show the significance of each slope. By looking at the independent relationships with p-values less than 0.05, we can see that sugars and carbohydrates significantly affect the amount of calories in cereal with a positive slope of 3.25 and 3.08 respectively. Both of these variables have a p-value that is much smaller than 0.05. Next, we take a look at pairs of variables that affect the calorie content. The combination of fats and sugars as well as the combination of carbohydrates and proteins both have a p-value less than 0.05, making the data valid. That said, they have coefficients of 0.42 and 0.16 respectively, meaning that there is not as much significance compared to the independent slopes.

```
sugar.lm <- lm(calories ~ sugars, data = UScereal)
plot(calories ~ sugars, data = UScereal)
abline(sugar.lm, col = "green")
```

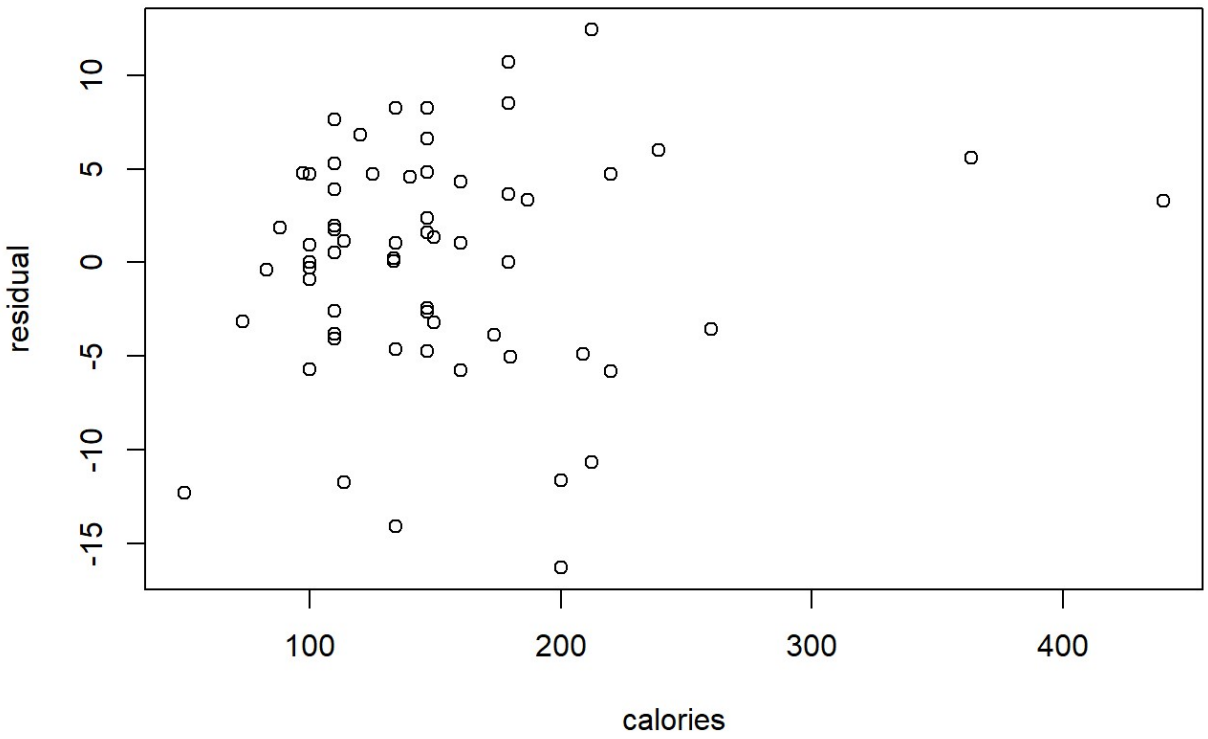



```
carbo.lm <- lm(calories ~ carbo, data = UScereal)
plot(calories ~ carbo, data = UScereal)
abline(carbo.lm, col = "red")
```



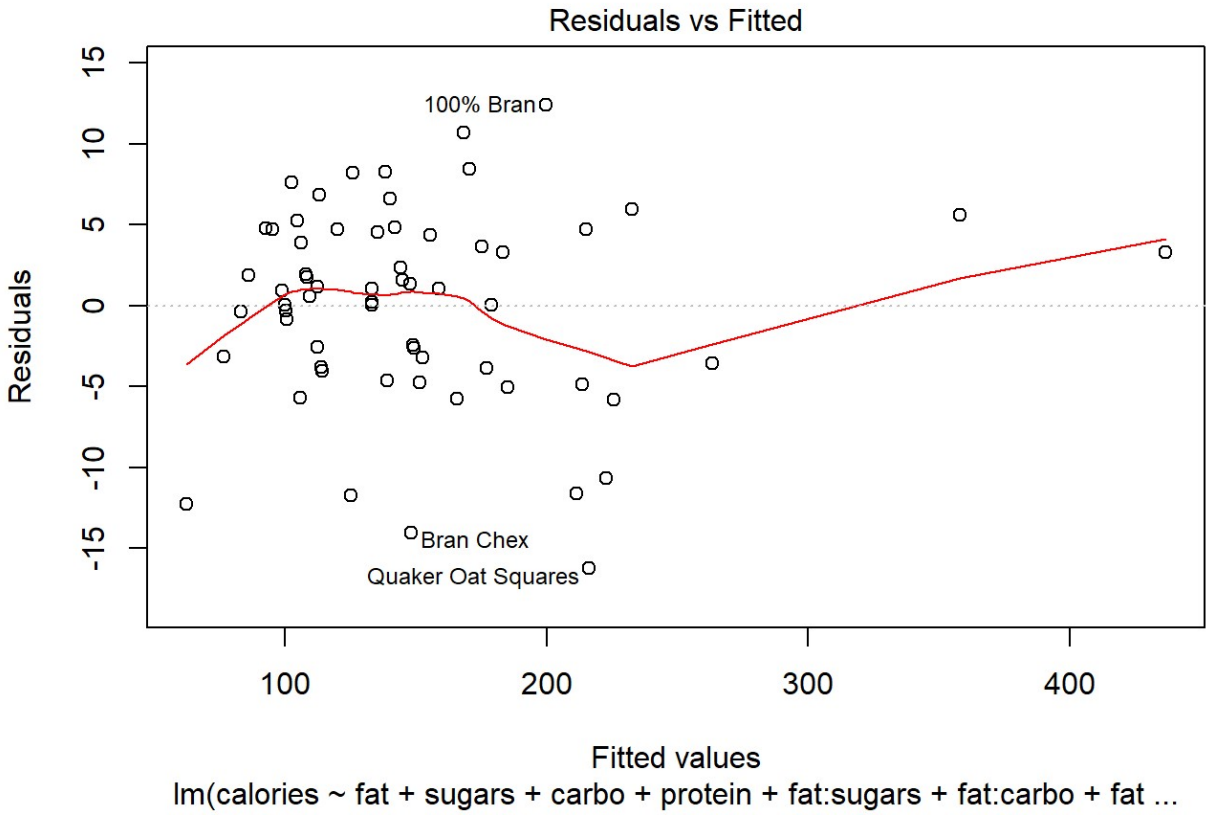
In the above section, I included a linear regression of the siginificant variables, sugar and carbohydrates, compared to the calories, which indeed gave us a positive slope.

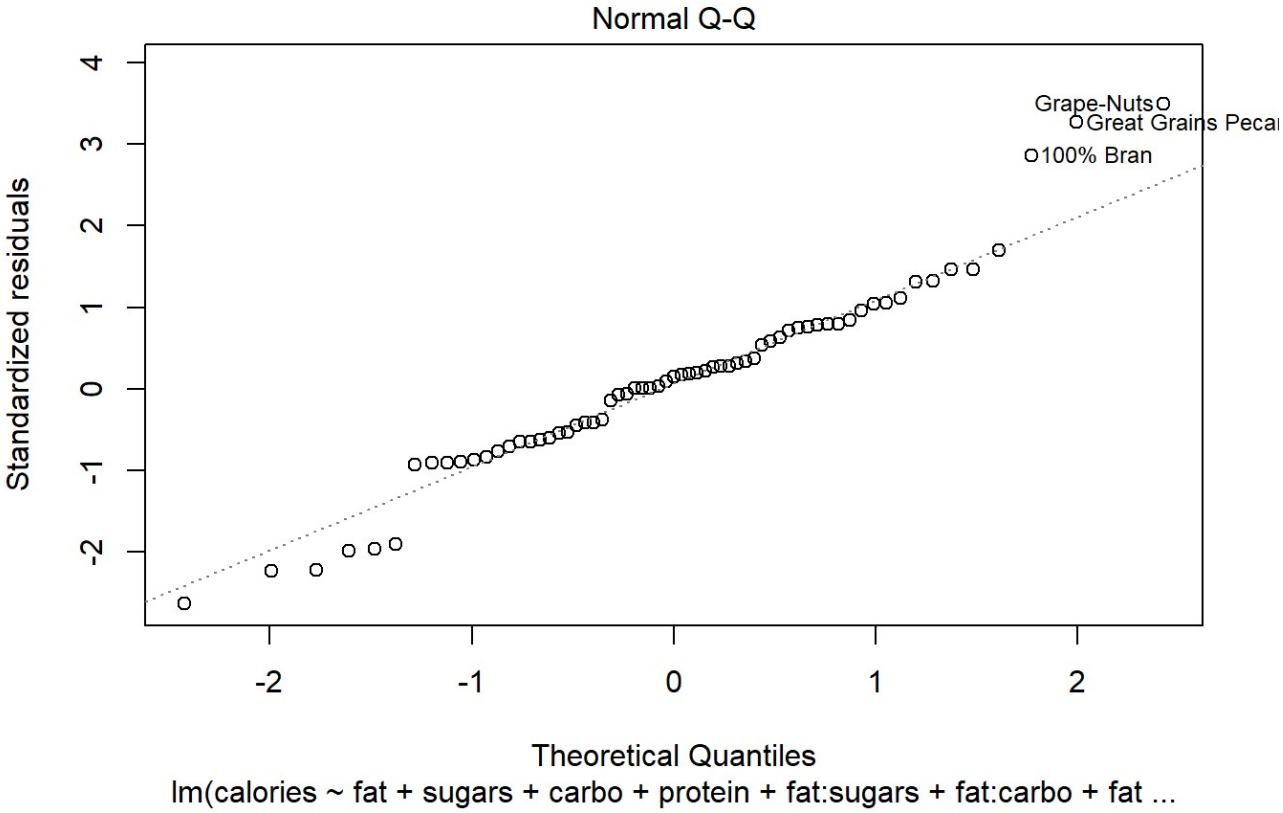
```
residual <- resid(cereal)
plot(residual ~ calories, data = UScereal)
```

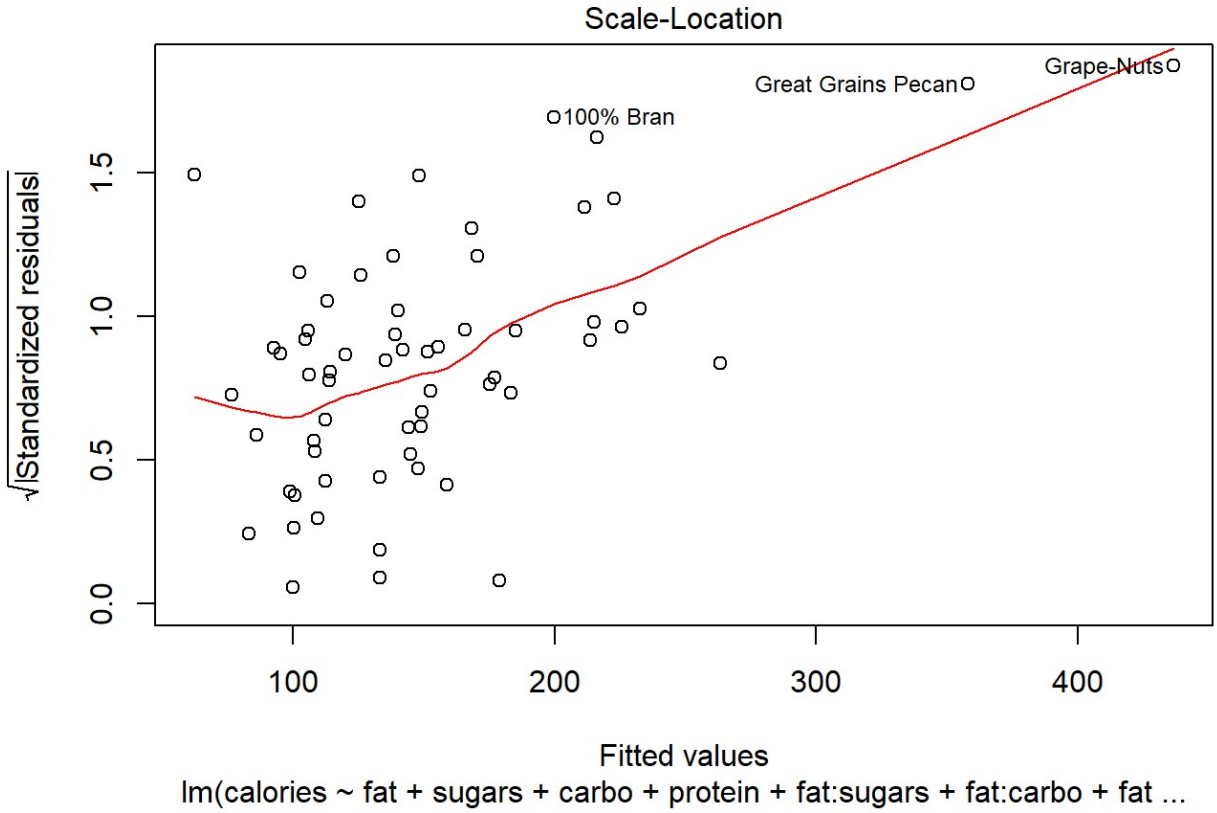


When we plot our residual compared to our calories, we can see that there is an increased density in the number of points in one area, suggesting that our linear regression model may have been inadequate in accurately predicting the results.

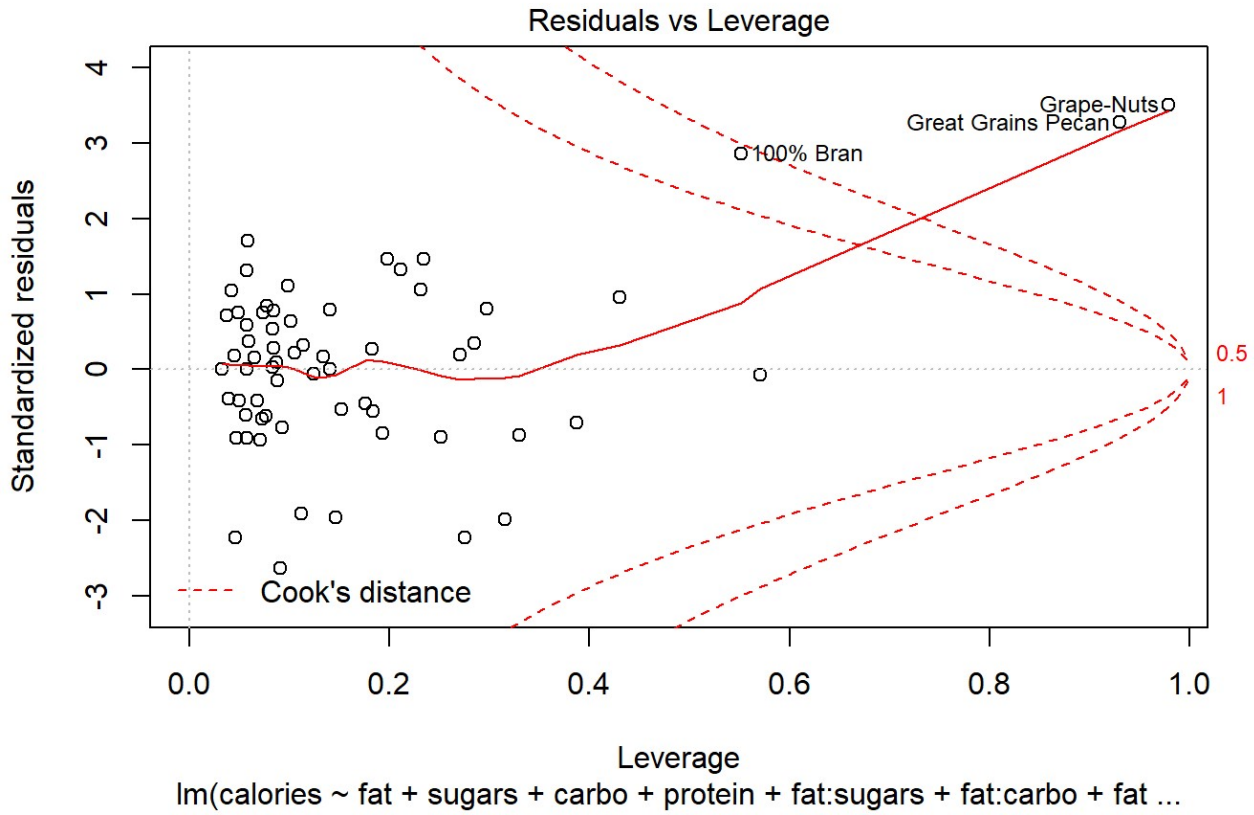
```
plot(cereal)
```


Residuals vs Fitted
Residuals
15
10
5
0
-5
-10
-15
100% Bran
Bran Chex
Quaker Oat Squares
100
200
300
400
Fitted values
lm(calories ~ fat + sugars + carbo + protein + fat:sugars + fat:carbo + fat ...


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(calories ~ fat + sugars + carbo + protein + fat:sugars + fat:carbo + fat ...
Grape-Nuts
Great Grains Pecan
100% Bran
-2
-1
0
1
2
3
4




```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



Finally, let's take a look at the quality control graphs from the multiple regression call. The first graph is the same as the residual graph that we looked at earlier. The second graph shows whether the residuals are normaly distributed. We can see some deviation from the line, especially on the edges. The third graph can tell whether the data is homeoscedastic. Because of the structure in the plot, we have heteroscedastic data. In the fourth plot, we see dashed lines, stating that there are outliers in the data. Given the four quality control graphs, we can assume that linear regression is not the best approach in determining our hypothesis for this data.

Finally, in response to the research question, calorie content is determined by the amount of molecules, although the residuals and quality control charts tell us that this data is not reliable.