Learning More About Cereals

A report submitted to Dr. Eliana Christou

Applied Statistics 1

Fall 2018

Author: Akshay D Patil

Abstract:

The goal of the project is to perform a thorough analysis of the dataset for different breakfast cereals available on the market by using R. The task is to build up a multiple linear regression model that predicts the rating of breakfast cereals according to its nutritional values and shelf location. The model is also examined for adequacy and treatment is carried out for outliers and missing values. Some interesting results were concluded at the end of our analysis which suggested that cereals with higher protein and fiber contents and lower contents in fats, sugars, and sodium tend to have high rating irrespective of the cereal's position on the shelf.

Table of Contents

1 Introduction:

The best way to start a healthy lifestyle is to start your day with a healthy breakfast. Almost half of the people around the world eat cereals for breakfast, without knowing for a fact that the cereals they are eating are healthy or not. But choosing a right cereal is not always easy. In this study, we will be developing a model that predicts a rating for different kinds of cereals based on certain variables by performing statistical analysis on a dataset.

2 Data Description:

The dataset contains per-serving nutritional records of 77 different kinds of cereals from 7 different manufacturers. To create simple random sample workers have been randomly sent into grocery stores and have collected data concerning the nutritional value and shelf location for 77 cereals, so that each cereal in the full Dataset has an equal probability of being selected.

Consumer Reports rating is included within the dataset, that is calculated by means of an undisclosed system possibly primarily based at the dietary content material. There are 7 nutrients in overall, which encompass protein, fats, sodium, fiber, carbohydrates, potassium (potass), and vitamins. Calories per serving and sugars are also taken into consideration as nutritional content. Other statistics we have are the manufacturer (mfr), supermarket show shelf place (shelf), type of cereal whether it's hot or cold (type), recommended serving size in ounces (weight), and recommended serving size in cups. No other exterior data is used while developing the model.

2.1 Treatment for Missing values:

After analyzing the dataset, it is noted that three kinds of cereals, Almond Delight, Cream of Wheat and Quaker Oatmeal, have data missing for the variables potass, carbo and sugars which is recorded as '-1'. These Observations are excluded from our analysis. So, we will be performing our analysis on 74 remaining cereals.

2.2 Variable Description:

2.2.1 Quantitative Variables:

In the Dataset, the number of calories per serving, grams of protein, grams of fat, milligrams of sodium, grams of

fiber, grams of carbohydrates, grams of sugars, milligrams of potassium, the weight of one serving and the number of cups in one serving are the quantitative variables.

2.2.2 Categorical Variables:

Categorical variables are those variables that take on usually fixed number of possible values; that is, they assign each observation to a particular group. They are also known as qualitative variables. In our dataset cereal manufacturer, type, typical percentage of the FDA's RDA of vitamins and the shelf location are the categorical variables.

2.2.3 Descriptive Statistics of Independent and Categorical variables:

Quantitative Variables	Mean	Median	Std. Deviation	Min	Max
Calories	107	110	19.844	50	160
Protein (grams)	2.514	2.5	1.076	1	6
Fat (grams)	1	1	1.007	0	5
Sodium (mg)	162.4	180	82.77	0	320
Fiber (grams)	2.176	2	2.423	0	14
Carbohydrates (grams)	14.73	14.50	3.892	5	23
Sugar (grams)	7.108	7	4.36	0	15
Potassium (mg)	98.51	90	70.88	15	330
Weight (per serving)	1.031	1	0.153	0.5	1.5
Cups	0.8216	0.75	0.236	0.25	1.5

Cups denotes number of cups per serving

Manufacturer:

MFR	A	G	K	N	P	Q	R
Frequency	1	22	23	5	9	7	7

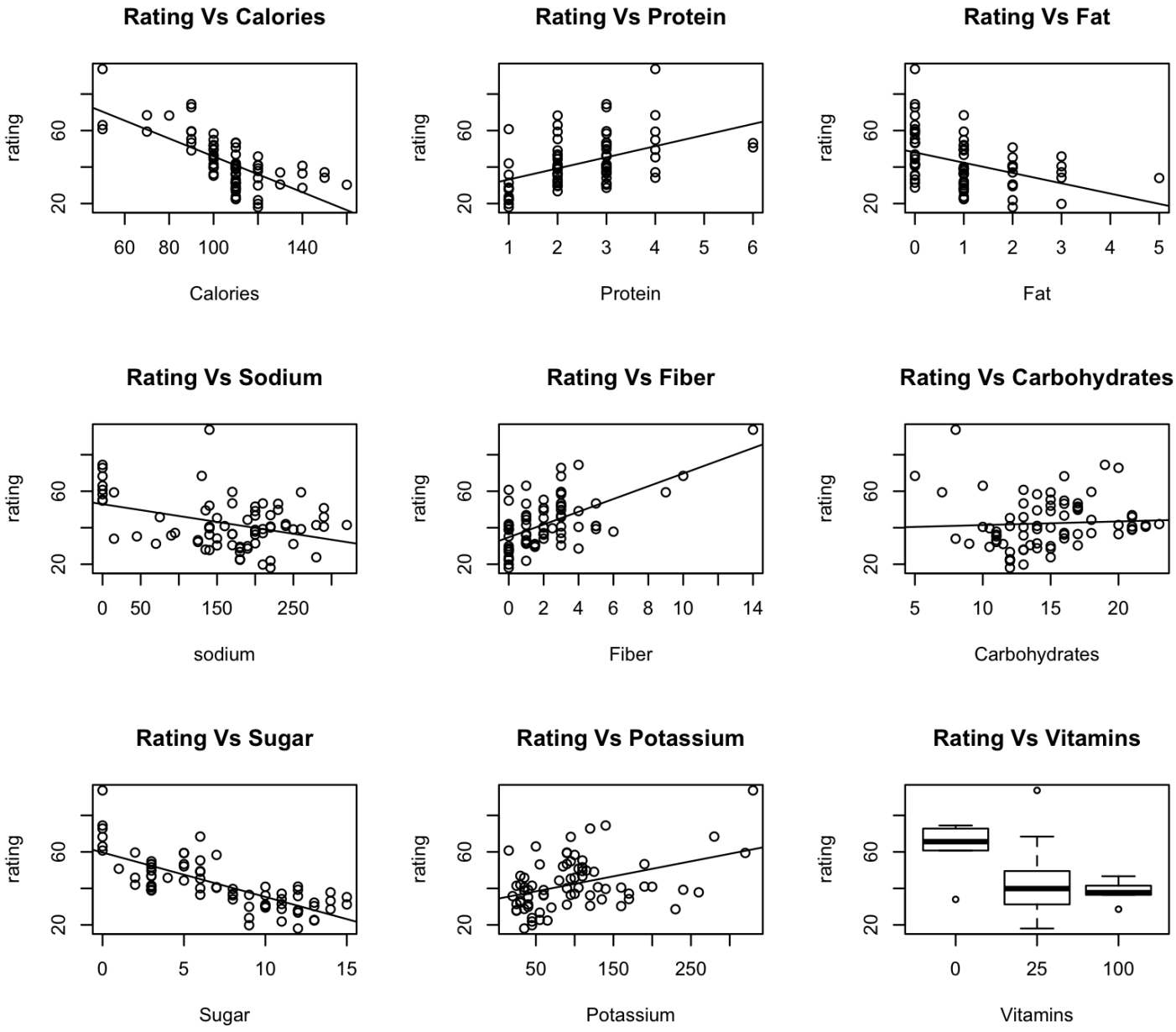
Shelf: **Vitamins:** **Type:**

Shelf	Frequency		Vitamins	Frequency		Type	Frequency
1	19		0	6		Hot	73
2	20		25	62		Cold	1
3	35		100	6		Total	74

Vitamins - Typical percentage of the FDA's RDA of vitamins (0, 25, or 100

2.3 Data Analysis

Ratings of cereals are usually calculated based on the nutritional content. So, let us first examine the relationship between dependent variable rating and the 9 independent variables: 7 nutrients which are protein, fat, sodium, fiber, carbohydrates, potassium, vitamins and also calories and sugars.



Scatter plots 1

2.4 Correlation Matrix Between the Variables

Rating	**Rating**								
Potassium	0.42	**Potassium**							
Sugar	-0.76	0.02	**Sugar**						
Carbs	0.05	-0.37	-0.45	**Carbs**					
Fiber	0.58	0.91	-0.14	-0.38	**Fiber**				
Sodium	-0.4	-0.039	0.04	0.3	-0.07	**Sodium**			
Fat	-0.41	0.2	0.28	-0.28	0.014	-0.01	**Fat**		
Protein	0.47	0.57	-0.28	-0.03	0.5	0.012	0.21	**Protein**	
Calories	-0.69	0.07	0.55	0.27	-0.29	0.29	0.50	0.034	**Calories**

From scatter plots and the correlation matrix we can conclude that variables like calories, sugars, sodium, and vitamins are negatively correlated with the rating while variables like potassium, fiber and protein are positively correlated with the rating. We can also see, fiber has the strongest positive correlation with the rating, while sugar has the most negative correlation with the rating, followed by calories. Also, the correlation between rating and carbohydrates is very low, which means that rating is weakly dependent on carbohydrate. Scatter plot also suggests some outliers for which we will have to work on to make a good model.

3 Model Description

3.1 Mathematical Notion of the Model

Multiple Regression model is given as

$$\boldsymbol{y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon_i}$$

where $\boldsymbol{y_i}$ is the response variable or the predicted variable

and $\boldsymbol{x_1, x_2, x_3}$ are the predictor variables or the independent variables

where as $\boldsymbol{\beta_0}$, $\boldsymbol{\beta_1, \beta_2, \beta_3}$ are the regression coefficients and $\boldsymbol{\varepsilon_i}$ are the uncorrelated errors with mean zero and constant variance $\boldsymbol{\sigma^2}$.

Assumptions for a good working model:

- Residuals must be normally distributed.
- A linear relationship is assumed between dependent and independent variables.
- The residuals are homoscedastic i.e. constant variance.
- Independent variables should not be highly correlated with each other.

Multiple linear regression is used to calculate point estimates. It tells us how much the dependent variable will change with a unit change in an independent variable.

3.2 Purpose of the Model:

The task is to build up a model that predicts the rating of breakfast cereals according to its nutritional values and other available independent variables. Since there are more than one independent variable Multiple Regression Model is the best fit for our problem. In the end, the model will help us in calculating the ratings of particular cereal given its nutritional content.

3.3 Model development

The model used for predicting the ratings of the cereals is Multiple Linear Regression Model. After selecting the model, the next important phase was to select the independent variables which will be used in the model. Ratings of cereals are usually calculated based on the nutritional content. So, a model was build using rating as the predicted variable and all 9 nutrients as independent variables. The R^2 of the model was 100% which suggested

that there are variables present which are highly correlated which other. Considering the correlation matrix, potassium was dropped from the model since it is highly correlated with fiber. Also, calories are highly correlated with fat and sugar. So, calories is dropped from the model too and the model is fitted again. Running this model, we found that carbohydrates is insignificantly related to ratings. The correlation matrix suggests that vitamins are weakly related to ratings, so vitamins are also dropped from the model along with carbs. Dropping carbs and vitamins did not make any significant change in R^2 since they were insignificantly related to the dependent variable. Since dropping calories from the model, carbs became insignificant, an interaction term of carbs and calories was incorporated in the model but that too was found insignificant. After all this elimination, R^2 was hardly affected and was still 99% suggesting, the dropped variables were correlated to the other independent variables. So, now out of the remaining five independent variables, each variable was dropped one by one and the R^2 value was noted down. It was concluded that dropping protein did not affect the R^2 value, this can be backed up by the fact from correlation matrix that protein and fibers are strongly related. Since, protein was a significant variable, an interaction term of protein and fibers was incorporated and the model was fitted again. This model had an R^2 value of 97% and also all the variables were significantly related to the dependent variable rating even at 1%. So, thus the final model had three independent variables (fats, sodium, sugars) and an interaction term of protein and fibers. The final model:

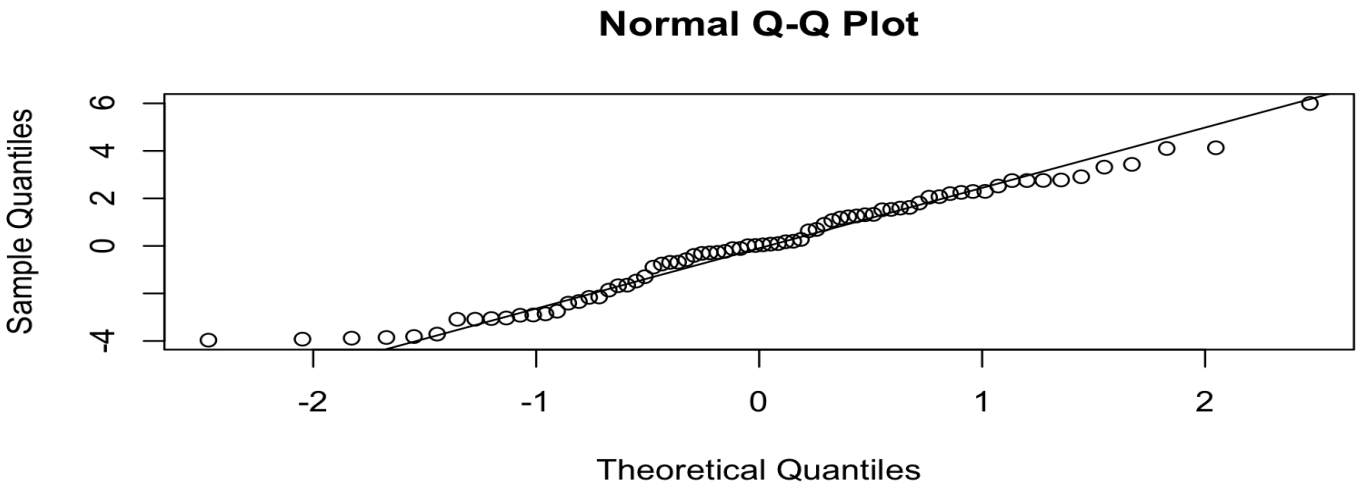
$$\boldsymbol{rate = \beta_0 + \beta_1 sug + \beta_2 sod + \beta_3 fat + \beta_4 profib + \varepsilon}$$

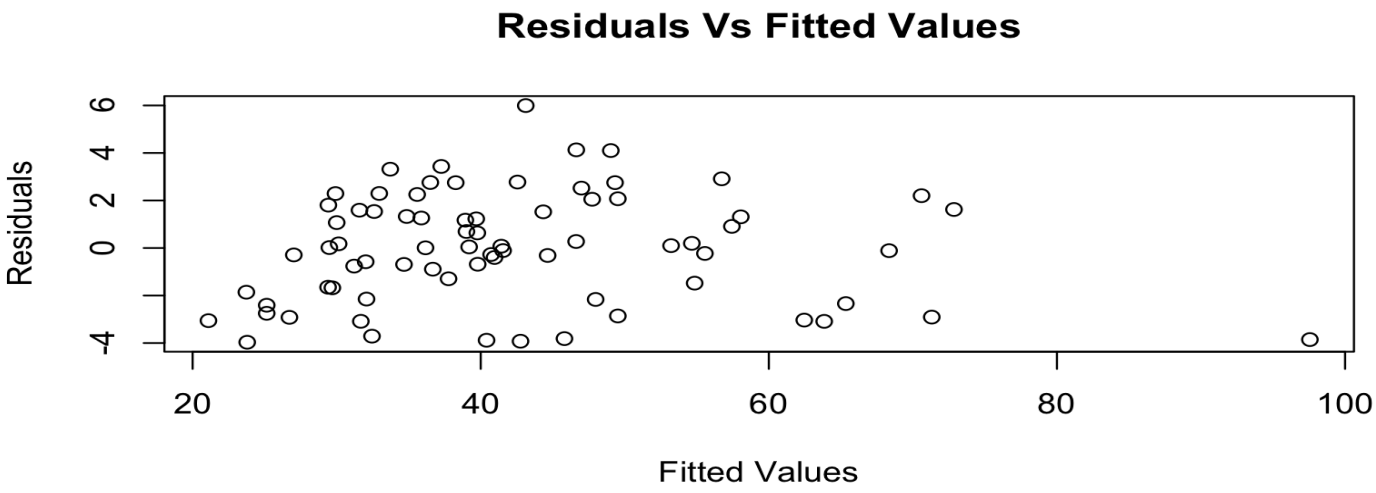
where sug represents sugar, sod represents sodium, fat represents fats and profib represents an interaction term of protein and fiber.

3.4 Model Adequacy Checking:

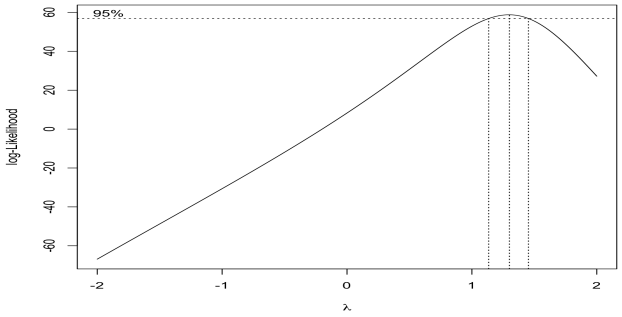
After fitting the model, it is important to investigate the residuals to determine any outliers and whether or not they appear to fit the assumption of a normal distribution. Thus, a normal quantile plot is plotted and also residuals

are plotted against the fitted values of rating.



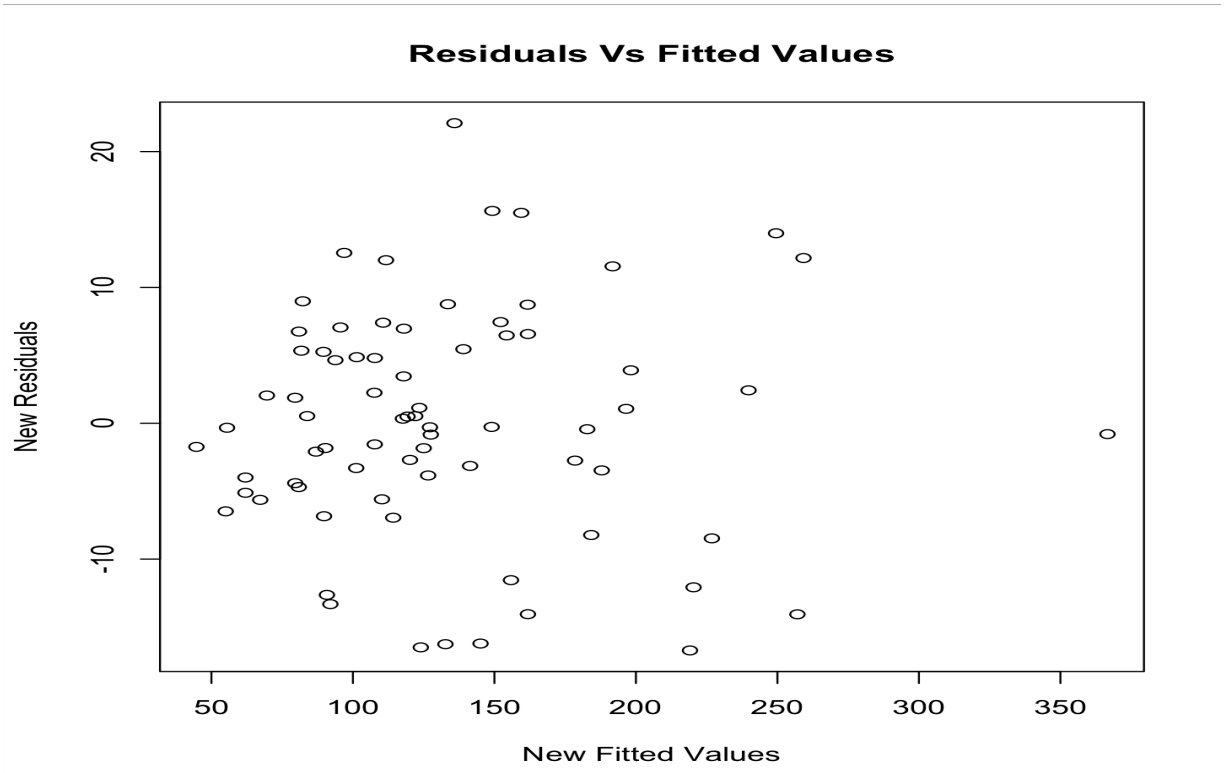


From the first plot, we observe that the residuals lie on almost a straight line, moreover, we can also observe some possible outliers, which needs further investigation. We observe that the residuals look randomly scattered, with no pattern, around zero. No violation from the assumptions is visible. Moreover, we observe some possible outliers present, and they needed more investigation to check whether they are influential points or not. We also run the Shapiro-Wilk test, and find p value to be about 12%, which is not high. So, we carry out the box cox transformation.



Box-Cox Transformation test

We can note that the value of λ =1.3, which suggests that our model needs a transformation to stabilize the variance. After carrying out the transformations, we again plot the normal quantile plot and the fitted values vs Residuals plot.



We observe that the residuals look randomly scattered with no pattern around zero. Also, the Shapiro-Wilk test, gives p value = 51% which is pretty high and hence we conclude that normality assumption is satisfied.

3.5 Investigating and treatment of Outliers:

From the normal quantile plot, we did observe some outliers. They were then checked if those outliers were influential points or not. DFIT test was carried out to check for influential point and observation 1, 3, 4, 11, 12 were found out to be influential. The model was then refitted again by excluding these observations one by one and as a group and R^2 and MSres was noted. There were no significant changes in MSres and R^2. Thus, it was concluded that no outliers were influential and hence no observation was deleted.

3.6 Hypothesis testing

All the variables used are found out be significant at 1 % level of significance, which shows that our model is working fine. Also, the overall F-test, suggests that at least one of the predictor is significant to the dependent variable.

4 Model Output:

```
> newfinalmodel=lm(rating ~ fat + sod + profib + sug)
> summary(newfinalmodel)

Call:
lm(formula = rating ~ fat + sod + profib + sug)

Residuals:
     Min       1Q   Median       3Q      Max
-16.7261  -4.6259  -0.2827   5.4294  22.0954

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 220.3681     2.9602   74.44   <2e-16 ***
fat         -14.8677     1.0501  -14.16   <2e-16 ***
sod          -0.2505     0.0122  -20.54   <2e-16 ***
profib        3.2383     0.1112   29.13   <2e-16 ***
sug          -7.5668     0.2462  -30.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.617 on 69 degrees of freedom
Multiple R-squared:  0.9792,  Adjusted R-squared:  0.978
F-statistic: 813.7 on 4 and 69 DF,  p-value: < 2.2e-16

> anova(newfinalmodel)
Analysis of Variance Table

Response: rating
          Df Sum Sq Mean Sq F value    Pr(>F)
fat        1  39973   39973  538.35 < 2.2e-16 ***
sod        1  37242   37242  501.57 < 2.2e-16 ***
profib     1  94359   94359 1270.81 < 2.2e-16 ***
sug        1  70113   70113  944.27 < 2.2e-16 ***
Residuals 69   5123      74
---
```

The prediction equation of the regression line is given as

$$rating = 220.3681 - 14.8677\ fat - 0.2505\ sodium - 7.5668\ sugar + 3.2383\ profib$$

4.1 Interpretation of coefficients:

- **Fat**

The variable fat is negatively related with the dependent variable rating, which means as the fat contents in cereal increases, the rating decreases. With one unit change in fat content in cereals, all else zero, the rating of cereal decreases by 15.

- **Sodium**

The variable sodium is negatively related with the dependent variable rating, which means as the sodium contents in cereal increases, the rating decreases. With one unit change in sodium content in cereals, all else zero, the rating of cereal decreases by 0.25.

- **Sugar**

The variable sugar is also negatively related with the dependent variable rating, which means as the sugar contents in cereal increases, the rating decreases. With one unit change in sodium content in cereals, all else zero, the rating of cereal decreases by 7.6.

- **Profib (Protein and fiber)**

The variable profib is positively related with the dependent variable rating, which means as the protein and fibers contents in cereal increases, the rating increases. With one unit change in protein content in cereals, keeping other variables constant, expected rating increases by 3.2. Also With one unit change in fiber content in cereals, keeping other variables constant, expected rating increases by 3.2.

5 Investigation of other Variables

Although cereal rating, cannot be calculated based on variables apart from nutrients, we are still interested in examining the relationship between rating and variables like manufacturer, shelf position etc.

5.1 Rating vs Manufacturer

Let's examine the relationship between Manufacturer and Ratings of the Cereals.

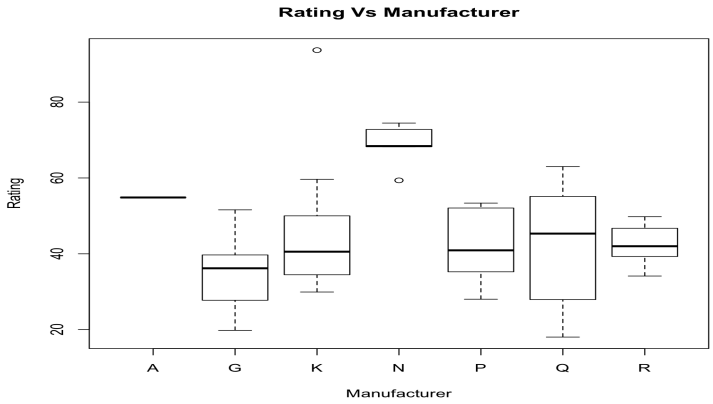
```
> ls=lm(rate~mfr)
> summary(ls)

Call:
lm(formula = rate ~ mfr)

Residuals:
    Min      1Q  Median      3Q     Max
-23.743  -7.705  -0.678   5.768  49.666

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    54.85      11.80   4.649 1.62e-05 ***
mfrG          -20.37      12.06  -1.688   0.0961 .
mfrK          -10.81      12.05  -0.897   0.3729
mfrN           13.80      12.93   1.068   0.2894
mfrP          -13.15      12.44  -1.057   0.2944
mfrQ          -13.07      12.61  -1.036   0.3040
mfrR          -12.29      12.61  -0.974   0.3336
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.8 on 67 degrees of freedom
Multiple R-squared:  0.3512,  Adjusted R-squared:  0.2931
F-statistic: 6.043 on 6 and 67 DF,  p-value: 4.276e-05
```



Manufacturer A is our reference group. From the above plot and the regression results it seems that Rating is related to who the manufacturer is. From the plot, it seems manufacturer 'N' tends to produce highest rated cereals as compared to other manufacturers. Cereals manufactured by manufacturer 'G' has rating by 20 than manufacture A. Similar interpretation can be done for other manufacturer.

5.2 Rating vs Shelf position

Let's examine the relationship between Manufacturer and Shelf Position of the Cereals.

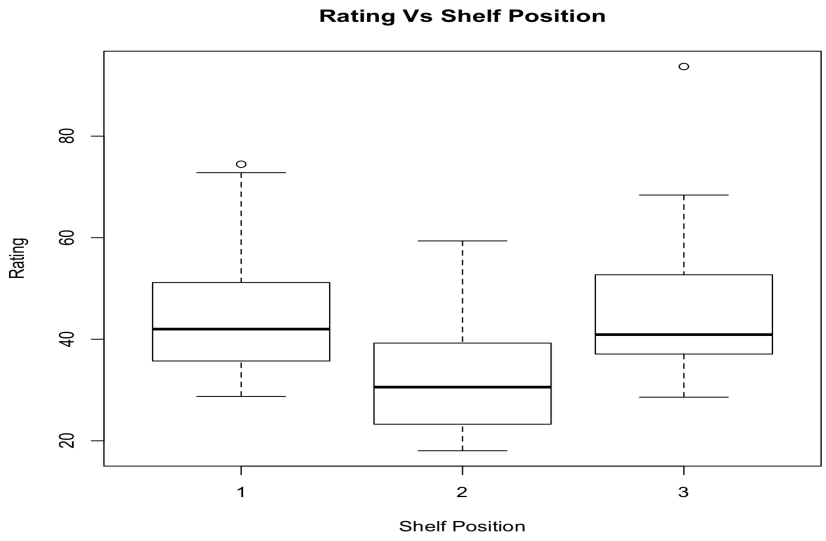
```
> ls=lm(rate~shlf)
> summary(ls)

Call:
lm(formula = rate ~ shlf)

Residuals:
    Min      1Q  Median      3Q     Max
-17.157  -9.045  -4.240   5.760  48.175

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  45.8990     3.0092  15.253  < 2e-16 ***
shlf2       -12.4042     4.2021  -2.952  0.00428 **
shlf3        -0.3694     3.7378  -0.099  0.92156
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.12 on 71 degrees of freedom
Multiple R-squared:  0.1503,  Adjusted R-squared:  0.1264
F-statistic: 6.281 on 2 and 71 DF,  p-value: 0.003078
```



The second plot suggests that there is a non-linear relation between the rating of the cereal and shelf position of the cereal. The data does not follow any pattern. The highest rated cereal is on the level three. Also, there are other low rated cereals on the same level. Some high rated cereals are also on the lowest shelf. So, we can conclude that there is no specific pattern followed in terms of rating and shelf position of cereals. Also, it makes no sense to interpret the regression coefficients.

6 Conclusion

A multiple linear regression model was used to predict ratings of cereal. The independent variables used to build the model are statistically significant. Manufacturer N produces cereals with the highest rating than others. The mean rating of the manufacturer 'G' is the lowest. Manufacturer N produces cereals with the least sugar content while manufacturer 'P' produces cereals with highest sugar content as compared to manufacturer 'A'.

The highest rated cereal is placed on the third level, the one closest to eye level based on normal store shelves. But that doesn't conclude that the highest rated cereals are on level three. So, we can conclude that instead of placing the cereals on shelves according to their rating they might be placed according to their popularity and also where one can easily find them. Some high rated cereals are also placed on level 1, usually where cheaper cereals are placed. This makes us believe that there might be other factors like the cost of the cereal that might be another significant variable related to rating which is not considered in this dataset. Furthermore, it makes sense that rating cannot be based on shelf position, as if higher the position of cereals on shelves, higher their ratings!

The cereal rating is highly correlated with fibers and protein and also highly and negatively correlated with sugars and fat. This concludes that cereal with the higher rating has higher protein and fiber content and lower sugar and fat contents. This can be backed by our data which shows that the lowest rated cereal Cap'n'Crunch which has the lowest rating has high fat and high sugar content while low protein content and have no fibers at all. Also, the highest rated cereals All-Bran with Extra Fiber has highest fiber content and protein content and zero fat and zero sugars. Also, most of the cereals manufactured by 'N' have high fiber content and negligible fats and sugar content.

Thus, from our analysis, we can be sure that cereal with higher content in fibers and proteins and low content in fats and sugar can make up for healthy and nutritious breakfast, even though it has been picked from lowest shelf position!

7 Appendix

R Code from Next Page

APENDIX (continued)

```
# read the data after dealing with '-1' in the data set

cereals=read.table('cereal.txt',header=T)
summary(cereals)
```

```
                        name    mfr    type      calories      protein
 100%_Bran                : 1   A: 1   C:73   Min.   : 50   Min.   :1.000
 100%_Natural_Bran        : 1   G:22   H: 1   1st Qu.:100   1st Qu.:2.000
 All-Bran                 : 1   K:23          Median :110   Median :2.500
 All-Bran_with_Extra_Fiber: 1   N: 5          Mean   :107   Mean   :2.514
 Apple_Cinnamon_Cheerios  : 1   P: 9          3rd Qu.:110   3rd Qu.:3.000
 Apple_Jacks              : 1   Q: 7          Max.   :160   Max.   :6.000
 (Other)                  :68   R: 7
      fat        sodium          fiber            carbo           sugars
 Min.   :0   Min.   :  0.0   Min.   : 0.000   Min.   : 5.00   Min.   : 0.000
 1st Qu.:0   1st Qu.:135.0   1st Qu.: 0.250   1st Qu.:12.00   1st Qu.: 3.000
 Median :1   Median :180.0   Median : 2.000   Median :14.50   Median : 7.000
 Mean   :1   Mean   :162.4   Mean   : 2.176   Mean   :14.73   Mean   : 7.108
 3rd Qu.:1   3rd Qu.:217.5   3rd Qu.: 3.000   3rd Qu.:17.00   3rd Qu.:11.000
 Max.   :5   Max.   :320.0   Max.   :14.000   Max.   :23.00   Max.   :15.000

     potass          vitamins          shelf           weight           cups
 Min.   : 15.00   Min.   :  0.00   Min.   :1.000   Min.   :0.500   Min.   :0.2500
 1st Qu.: 41.25   1st Qu.: 25.00   1st Qu.:1.250   1st Qu.:1.000   1st Qu.:0.6700
 Median : 90.00   Median : 25.00   Median :2.000   Median :1.000   Median :0.7500
 Mean   : 98.51   Mean   : 29.05   Mean   :2.216   Mean   :1.031   Mean   :0.8216
 3rd Qu.:120.00   3rd Qu.: 25.00   3rd Qu.:3.000   3rd Qu.:1.000   3rd Qu.:1.0000
 Max.   :330.00   Max.   :100.00   Max.   :3.000   Max.   :1.500   Max.   :1.5000

     rating
 Min.   :18.04
 1st Qu.:32.45
 Median :40.25
 Mean   :42.37
 3rd Qu.:50.52
 Max.   :93.70
```

```
rate=cereals$rating
cal=cereals$calories
pro=cereals$protein
fat=cereals$fat
sod=cereals$sodium
fib=cereals$fiber
carb=cereals$carbo
sug=cereals$sugars
pot=cereals$potass
wgt=cereals$weight
cups=cereals$cups
```

```
mfr=cereals$mfr
typ=cereals$type
vit=cereals$vitamins
shlf=cereals$shelf

mfr=factor(mfr)
typ=factor(typ)
vit=factor(vit)
shlf=factor(shlf)
```

```
#Correlation between Variables
X=cbind(cal,pro,fat,sod,fib,carb,sug,pot,wgt,cups,vit)
cor(X)
```

```
             cal         pro           fat           sod         fib        carb
cal   1.00000000  0.03399166  0.5073732397  0.2962474981 -0.29521183  0.27060605
pro   0.03399166  1.00000000  0.2023533963  0.0115588913  0.51400610 -0.03674326
fat   0.50737324  0.20235340  1.0000000000  0.0008219036  0.01403587 -0.28493369
sod   0.29624750  0.01155889  0.0008219036  1.0000000000 -0.07073492  0.32840919
fib  -0.29521183  0.51400610  0.0140358654 -0.0707349230  1.00000000 -0.37908370
carb  0.27060605 -0.03674326 -0.2849336855  0.3284091857 -0.37908370  1.00000000
sug   0.56912054 -0.28658397  0.2871524866  0.0370589612 -0.15094850 -0.45206919
pot  -0.07136125  0.57874284  0.1996367171 -0.0394380876  0.91150392 -0.36500293
wgt   0.69645215  0.23067141  0.2217141647  0.3125335701  0.24629218  0.14480528
cups  0.08919615 -0.24209861 -0.1575787041  0.1195841083 -0.51369716  0.35828371
vit   0.37457928  0.06281245  0.0000000000  0.5041306084 -0.02788396  0.19968216
              sug          pot        wgt        cups          vit
cal   0.569120535 -0.071361247  0.6964521  0.08919615  0.374579284
pro  -0.286583967  0.578742837  0.2306714 -0.24209861  0.062812452
fat   0.287152487  0.199636717  0.2217142 -0.15757870  0.000000000
sod   0.037058961 -0.039438088  0.3125336  0.11958411  0.504130608
fib  -0.150948502  0.911503921  0.2462922 -0.51369716 -0.027883958
carb -0.452069189 -0.365002934  0.1448053  0.35828371  0.199682157
sug   1.000000000  0.001413982  0.4605471 -0.03243610  0.232525853
pot   0.001413982  1.000000000  0.4205615 -0.50168832  0.009533718
wgt   0.460547135  0.420561534  1.0000000 -0.20171465  0.433855287
cups -0.032436100 -0.501688318 -0.2017146  1.00000000  0.058768414
vit   0.232525853  0.009533718  0.4338553  0.05876841  1.000000000
```

```
#Regressing each variable against rating and plotting it against rating
par(mfrow=c(3,3))
ls1=lm(rate~cal)
plot(cal,rate, xlab='Calories', ylab='rating', main='Rating Vs Calories')
abline(ls1)

ls2=lm(rate~pro)
```

```
plot(pro,rate, xlab='Protein', ylab='rating', main='Rating Vs Protein')
abline(ls2)

ls3=lm(rate~fat)
plot(fat,rate, xlab='Fat', ylab='rating', main='Rating Vs Fat')
abline(ls3)

ls4=lm(rate~sod)
plot(sod,rate, xlab='sodium', ylab='rating', main='Rating Vs Sodium')
abline(ls4)

ls5=lm(rate~fib)
plot(fib,rate, xlab='Fiber', ylab='rating', main='Rating Vs Fiber')
abline(ls5)

ls6=lm(rate~carb)
plot(carb,rate, xlab='Carbohydrates', ylab='rating', main='Rating Vs Carbohydrates')
abline(ls6)

ls7=lm(rate~sug)
plot(sug,rate, xlab='Sugar', ylab='rating', main='Rating Vs Sugar')
abline(ls7)

ls8=lm(rate~pot)
plot(pot,rate, xlab='Potassium', ylab='rating', main='Rating Vs Potassium')
abline(ls8)

ls9=lm(rate~vit)
plot(vit,rate, xlab='Vitamins', ylab='rating', main='Rating Vs Vitamins')
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------


#Model including all varaibles
ls=lm(rate~cal+pro+fat+sod+fib+carb+sug+pot+vit)
summary(ls)

Call:
lm(formula = rate ~ cal + pro + fat + sod + fib + carb + sug +
    pot + vit)

Residuals:
       Min         1Q     Median         3Q        Max
-5.371e-07 -2.319e-07  5.440e-08  2.005e-07  5.482e-07

Coefficients:
              Estimate Std. Error   t value Pr(>|t|)
(Intercept)  5.493e+01  2.806e-07 195753120   <2e-16 ***
cal         -2.227e-01  7.470e-09 -29817548   <2e-16 ***
pro          3.273e+00  5.675e-08  57677906   <2e-16 ***
fat         -1.691e+00  8.070e-08 -20959857   <2e-16 ***
sod         -5.449e-02  5.846e-10 -93213610   <2e-16 ***
```

```
fib          3.443e+00  4.882e-08  70530139   <2e-16 ***
carb         1.092e+00  3.479e-08  31396934   <2e-16 ***
sug         -7.249e-01  3.374e-08 -21485716   <2e-16 ***
pot         -3.399e-02  1.649e-09 -20620206   <2e-16 ***
vit25       -1.280e+00  1.935e-07  -6617571   <2e-16 ***
vit100      -5.121e+00  2.368e-07 -21627557   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.056e-07 on 63 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:      1
F-statistic: 1.539e+16 on 10 and 63 DF,  p-value: < 2.2e-16
--------------------------------------------------------------------------------



#Nutritional rating without potassium because of high correlation wth fibre
ls=lm(rate~cal+pro+fat+sod+fib+carb+sug+wgt+cups+vit)
Call:
lm(formula = rate ~ cal + pro + fat + sod + fib + carb + sug +
    wgt + cups + vit)

Residuals:
     Min       1Q   Median       3Q      Max
-1.69699 -0.36285  0.02578  0.47581  1.41482

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 56.351606   0.856534  65.790  < 2e-16 ***
cal         -0.185662   0.019037  -9.753 3.89e-14 ***
pro          2.828173   0.133942  21.115  < 2e-16 ***
fat         -2.238584   0.192562 -11.625  < 2e-16 ***
sod         -0.055456   0.001485 -37.349  < 2e-16 ***
fib          2.647560   0.071886  36.830  < 2e-16 ***
carb         0.949793   0.091256  10.408 3.11e-15 ***
sug         -0.920882   0.086907 -10.596 1.52e-15 ***
wgt         -2.817298   1.365222  -2.064   0.0432 *
cups         0.293730   0.492261   0.597   0.5529
vit25       -0.222109   0.478471  -0.464   0.6441
vit100      -4.080496   0.585238  -6.972 2.40e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7668 on 62 degrees of freedom
Multiple R-squared:  0.9975,    Adjusted R-squared:  0.997
F-statistic:  2217 on 11 and 62 DF,  p-value: < 2.2e-16
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------


#dropped weight (as it was highly correlated with calories) and also cups as it was not
 significant
```

```
ls=lm(rate~cal+pro+fat+sod+fib+carb+sug+vit)
Call:
lm(formula = rate ~ cal + pro + fat + sod + fib + carb + sug +
    vit)

Residuals:
     Min       1Q   Median       3Q      Max
-2.01082 -0.32910  0.03671  0.47396  1.54917

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 56.209704   0.705247  79.702  < 2e-16 ***
cal         -0.196228   0.018966 -10.346 2.71e-15 ***
pro          2.797485   0.133641  20.933  < 2e-16 ***
fat         -2.216331   0.197379 -11.229  < 2e-16 ***
sod         -0.055571   0.001501 -37.028  < 2e-16 ***
fib          2.528977   0.052619  48.062  < 2e-16 ***
carb         0.904934   0.086567  10.454 1.78e-15 ***
sug         -0.970396   0.081367 -11.926  < 2e-16 ***
vit25       -0.266503   0.482303  -0.553    0.582
vit100      -4.183957   0.598984  -6.985 1.98e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7878 on 64 degrees of freedom
Multiple R-squared:  0.9972,    Adjusted R-squared:  0.9968
F-statistic:  2567 on 9 and 64 DF,  p-value: < 2.2e-16
#no change in R2 seen is last model but there seems to be multiocolinearity present due
 to high R2 for above model
-------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------



#Since cal is highly corerelated with fat and also sugar lets chheck droppin cal
ls=lm(rate~pro+fat+sod+fib+carb+sug+vit)

Call:
lm(formula = rate ~ pro + fat + sod + fib + carb + sug + vit)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2131 -0.7205 -0.0669  0.8417  3.1258

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 56.522144   1.142992  49.451  < 2e-16 ***
pro          2.053149   0.182698  11.238  < 2e-16 ***
fat         -3.946465   0.170100 -23.201  < 2e-16 ***
sod         -0.055098   0.002433 -22.642  < 2e-16 ***
fib          2.478077   0.084985  29.159  < 2e-16 ***
carb         0.089382   0.058048   1.540    0.128
```

```
sug         -1.745491   0.051512 -33.885  < 2e-16 ***
vit25       -0.448055   0.781866  -0.573    0.569
vit100      -4.243538   0.971617  -4.368 4.61e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.278 on 65 degrees of freedom
Multiple R-squared:  0.9926,    Adjusted R-squared:  0.9917
F-statistic:  1092 on 8 and 65 DF,  p-value: < 2.2e-16

#droppinfg cal there was no change in R2 and carbs beacame insignificant,
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------



#So lets try dropping carbs from the model
ls=lm(rate~pro+fat+sod+fib+sug+vit)
summary(ls)
Call:
lm(formula = rate ~ pro + fat + sod + fib + sug + vit)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1775 -0.6993 -0.0434  0.8009  3.2142

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 57.979005   0.647912  89.486  < 2e-16 ***
pro          2.080005   0.183743  11.320  < 2e-16 ***
fat         -3.998973   0.168369 -23.751  < 2e-16 ***
sod         -0.054068   0.002364 -22.873  < 2e-16 ***
fib          2.410814   0.073651  32.733  < 2e-16 ***
sug         -1.784092   0.045465 -39.241  < 2e-16 ***
vit25       -0.359129   0.787787  -0.456 0.649980
vit100      -3.951282   0.962744  -4.104 0.000114 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.291 on 66 degrees of freedom
Multiple R-squared:  0.9923,    Adjusted R-squared:  0.9915
F-statistic:  1223 on 7 and 66 DF,  p-value: < 2.2e-16#R2 did not change by much
 therefore we drop carbs .
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------

#Also lets try dropping vit and check the R2
ls=lm(rate~pro+fat+sod+fib+sug)

Call:
lm(formula = rate ~ pro + fat + sod + fib + sug)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-5.3130 -0.7837 -0.0022  1.0135  3.4643

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 57.905645   0.753290  76.870  < 2e-16 ***
pro          1.985523   0.224875   8.829 6.86e-13 ***
fat         -3.931148   0.208486 -18.856  < 2e-16 ***
sod         -0.056653   0.002306 -24.568  < 2e-16 ***
fib          2.441897   0.092309  26.453  < 2e-16 ***
sug         -1.787778   0.048980 -36.500  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.622 on 68 degrees of freedom
Multiple R-squared:  0.9876,    Adjusted R-squared:  0.9866
F-statistic:  1079 on 5 and 68 DF,  p-value: < 2.2e-16#No significant difference in R2
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------

Since dropping calories from the model, carbs became insignificant, I tried
 incorporating interaction term of carbs and calories , but that too was found
 insignificant.
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------

#R2 value seems to be very high, there might be multicollinearity present
#lets try dropping one of these four variables one by one and lets check the R2 value
#it was concluded that dropping pro did not affect the R2 and hence protein is dropped
 from the final model and We can see that protein is 50% correlated to fibres. Thus we
 can see it is better to add an interaction term of protien and fibre and refir the
 model
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------



# fit the multiple linear regression
finalmodel=lm(rate ~ fat + sod + profib + sug)
summary(finalmodel)

Call:
lm(formula = rate ~ fat + sod + profib + sug)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9660 -1.8138  0.0319  1.6137  5.9942

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 63.843404   0.813719   78.46   <2e-16 ***
```

```
fat         -3.547086   0.288643  -12.29   <2e-16 ***
sod         -0.059458   0.003352  -17.74   <2e-16 ***
profib       0.750699   0.030562   24.56   <2e-16 ***
sug         -1.880715   0.067688  -27.79   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.369 on 69 degrees of freedom
Multiple R-squared:  0.9731,    Adjusted R-squared:  0.9715
F-statistic: 623.4 on 4 and 69 DF,  p-value: < 2.2e-16residuals=finalmodel$res
fittedvalues=finalmodel$fitted

-------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------


 # investigating for outliers
 par(mfrow=c(1,2))
qqnorm(residuals)
qqline(residuals)
plot(fittedvalues, residuals, xlab='Fitted Values', ylab='Residuals', main='Residuals
 Vs Fitted Values')
X=cbind(1, fat,  sod,  profib,  sug)
n=length(rate)
p=dim(X)[2]

-------------------------------------------------------------------------------------


#Identify outliers

MSres=(summary(ls)$sigma)^2
stand=res/sqrt(MSres)
which(abs(stand)>3)
#Nothing greater than 3 was found

# calculate leverage and measures of influence
#leverage
hat=X%*%solve(t(X)%*%X)%*%t(X)
lev=diag(hat)
which(lev>2*p/n)
# this were found leverage points [1] 1 2 3 4

# Checking outlier and influential points with cook's distance
# formula
r=residuals/sqrt(MSres*(1-lev))
cooks=(r^2/p)*(lev/(1-lev))

# command
cooks.distance(ls)
# cut-off point
which(cooks>1)
```

```
#nothing found

# Checking influential points DFFITS

# command
dffits(finalmodel)
# cut-off point
2*sqrt(p/n) # for DFFITS
which(abs(dffits)>2*sqrt(p/n))
# 1  3  4 11 12 this were found to be influential
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------


# Refit the model
finalmodel1=lm(rate[-1] ~ fat[-1] + sod[-1] + profib[-1] + sug[-1])
finalmodel2=lm(rate[-3] ~ fat[-3] + sod[-3] + profib[-3] + sug[-3])
finalmodel3=lm(rate[-4] ~ fat[-4] + sod[-4] + profib[-4] + sug[-4])
finalmodel4=lm(rate[-11] ~ fat[-11] + sod[-11] + profib[-11] + sug[-11])
finalmodel5=lm(rate[-12] ~ fat[-12] + sod[-12] + profib[-12] + sug[-12])
finalmodel6=lm(rate[-c(1,3,4,11,12)] ~ fat[-c(1,3,4,11,12)] + sod[-c(1,3,4,11,12)] +
 profib[-c(1,3,4,11,12)] + sug[-c(1,3,4,11,12)])


summary(finalmodel1)

Call:
lm(formula = rate[-1] ~ fat[-1] + sod[-1] + profib[-1] + sug[-1])

Residuals:
    Min      1Q  Median      3Q     Max
-4.8484 -1.6221  0.0445  1.6075  5.9350

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 63.765901   0.810453   78.68   <2e-16 ***
fat[-1]     -3.560077   0.286945  -12.41   <2e-16 ***
sod[-1]     -0.059655   0.003334  -17.89   <2e-16 ***
profib[-1]   0.770341   0.033552   22.96   <2e-16 ***
sug[-1]     -1.875379   0.067365  -27.84   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.353 on 68 degrees of freedom
Multiple R-squared:  0.9725,    Adjusted R-squared:  0.9709
F-statistic: 600.9 on 4 and 68 DF,  p-value: < 2.2e-16



summary(finalmodel2)
-------------------------------------------------------------------------------------------
Call:
```

```
lm(formula = rate[-3] ~ fat[-3] + sod[-3] + profib[-3] + sug[-3])

Residuals:
    Min      1Q  Median      3Q     Max
-4.7417 -1.5897  0.0221  1.6419  5.8961

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 63.65397    0.81882   77.74   <2e-16 ***
fat[-3]     -3.55366    0.28659  -12.40   <2e-16 ***
sod[-3]     -0.05870    0.00337  -17.42   <2e-16 ***
profib[-3]   0.76806    0.03272   23.47   <2e-16 ***
sug[-3]     -1.88000    0.06720  -27.98   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.352 on 68 degrees of freedom
Multiple R-squared:  0.9733,    Adjusted R-squared:  0.9717
F-statistic: 619.7 on 4 and 68 DF,  p-value: < 2.2e-16



summary(finalmodel3)
--------------------------------------------------------------------------------
Call:
lm(formula = rate[-4] ~ fat[-4] + sod[-4] + profib[-4] + sug[-4])

Residuals:
   Min     1Q Median     3Q    Max
-4.748 -1.509  0.085  1.642  5.837

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 63.761228   0.792858   80.42   <2e-16 ***
fat[-4]     -3.643687   0.284341  -12.81   <2e-16 ***
sod[-4]     -0.059615   0.003263  -18.27   <2e-16 ***
profib[-4]   0.802956   0.038065   21.09   <2e-16 ***
sug[-4]     -1.889200   0.065992  -28.63   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.305 on 68 degrees of freedom
Multiple R-squared:  0.9691,    Adjusted R-squared:  0.9673
F-statistic: 533.6 on 4 and 68 DF,  p-value: < 2.2e-16



summary(finalmodel4)
--------------------------------------------------------------------------------
Call:
lm(formula = rate[-11] ~ fat[-11] + sod[-11] + profib[-11] +
    sug[-11])
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.8372 -1.7888  0.1136  1.5875  6.1348

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 63.895305   0.799704   79.90   <2e-16 ***
fat[-11]    -3.642463   0.288017  -12.65   <2e-16 ***
sod[-11]    -0.060692   0.003357  -18.08   <2e-16 ***
profib[-11]  0.749181   0.030029   24.95   <2e-16 ***
sug[-11]    -1.853741   0.068017  -27.25   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.326 on 68 degrees of freedom
Multiple R-squared:  0.9743,    Adjusted R-squared:  0.9728
F-statistic: 643.8 on 4 and 68 DF,  p-value: < 2.2e-16



summary(finalmodel5)
-------------------------------------------------------------------------------------
Call:
lm(formula = rate[-12] ~ fat[-12] + sod[-12] + profib[-12] +
    sug[-12])

Residuals:
    Min      1Q  Median      3Q     Max
-4.0529 -1.7369 -0.0172  1.7485  5.9220

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 63.781763   0.802086   79.52   <2e-16 ***
fat[-12]    -3.422776   0.292753  -11.69   <2e-16 ***
sod[-12]    -0.059051   0.003309  -17.84   <2e-16 ***
profib[-12]  0.745220   0.030255   24.63   <2e-16 ***
sug[-12]    -1.885367   0.066709  -28.26   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.333 on 68 degrees of freedom
Multiple R-squared:  0.9733,    Adjusted R-squared:  0.9717
F-statistic: 619.9 on 4 and 68 DF,  p-value: < 2.2e-16

-------------------------------------------------------------------------------------
-------------------------------------------------------------------------------
#no point is influential and hence no need to delete observation
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------

BoxCox Transformation to for constant variance
```

```
install.packages('MASS')
library(MASS)
result=boxcox(rate ~ fat + sod + profib + sug,lambda=seq(from=-2, to=2, by=0.01))
result$x[result$y==max(result$y)]
[1] 1.3
#Result is 1.3 hence transformation is needed.
-------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------


#Transformed Model
newrate = (rate)^(1.3)
newfinalmodel=lm(newrate ~ fat + sod + profib + sug)

summary(newfinalmodel)

Call:
lm(formula = newrate ~ fat + sod + profib + sug)

Residuals:
     Min       1Q   Median       3Q      Max
-16.7261  -4.6259  -0.2827   5.4294  22.0954

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 220.3681     2.9602   74.44   <2e-16 ***
fat         -14.8677     1.0501  -14.16   <2e-16 ***
sod          -0.2505     0.0122  -20.54   <2e-16 ***
profib        3.2383     0.1112   29.13   <2e-16 ***
sug          -7.5668     0.2462  -30.73   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.617 on 69 degrees of freedom
Multiple R-squared:  0.9792,    Adjusted R-squared:  0.978
F-statistic: 813.7 on 4 and 69 DF,  p-value: < 2.2e-16

par(mfrow=c(2,1))
newresiduals=newfinalmodel$residuals
newfitted=newfinalmodel$fitted
> qqnorm(newresiduals)
> qqline(newresiduals)
> plot(newfitted, newresiduals, xlab='New Fitted Values', ylab='New Residuals',
 main='Residuals Vs Fitted Values')
> shapiro.test(newresiduals)

    Shapiro-Wilk normality test

data:  newresiduals
W = 0.98481, p-value = 0.5189

#Transformed model follows normal distribution
```

```
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------


# NEW FINAL MODEL
rating=newrate
newfinalmodel=lm(rating ~ fat + sod + profib + sug)
summary(newfinalmodel)

Call:
lm(formula = rating ~ fat + sod + profib + sug)

Residuals:
     Min       1Q   Median       3Q      Max
-16.7261  -4.6259  -0.2827   5.4294  22.0954

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 220.3681     2.9602   74.44   <2e-16 ***
fat         -14.8677     1.0501  -14.16   <2e-16 ***
sod          -0.2505     0.0122  -20.54   <2e-16 ***
profib        3.2383     0.1112   29.13   <2e-16 ***
sug          -7.5668     0.2462  -30.73   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.617 on 69 degrees of freedom
Multiple R-squared:  0.9792,    Adjusted R-squared:  0.978
F-statistic: 813.7 on 4 and 69 DF,  p-value: < 2.2e-16

anova(newfinalmodel)
Analysis of Variance Table

Response: rating
          Df Sum Sq Mean Sq F value    Pr(>F)
fat        1  39973   39973  538.35 < 2.2e-16 ***
sod        1  37242   37242  501.57 < 2.2e-16 ***
profib     1  94359   94359 1270.81 < 2.2e-16 ***
sug        1  70113   70113  944.27 < 2.2e-16 ***
Residuals 69   5123      74
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

--------------------------------------------------------------------------------
--------------------------------------------------------------------------------


# Investigating other variables
#rating vs manufacturer
ls=lm(rate~mfr)
summary(ls)
plot(mfr,rate, xlab='Manufacturer', ylab='Rating', main='Rating Vs Manufacturer')
```

```
Call:
lm(formula = rate ~ mfr)

Residuals:
    Min      1Q  Median      3Q     Max
-23.743  -7.705  -0.678   5.768  49.666

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    54.85      11.80   4.649 1.62e-05 ***
mfrG          -20.37      12.06  -1.688   0.0961 .
mfrK          -10.81      12.05  -0.897   0.3729
mfrN           13.80      12.93   1.068   0.2894
mfrP          -13.15      12.44  -1.057   0.2944
mfrQ          -13.07      12.61  -1.036   0.3040
mfrR          -12.29      12.61  -0.974   0.3336
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.8 on 67 degrees of freedom
Multiple R-squared:  0.3512,    Adjusted R-squared:  0.2931
F-statistic: 6.043 on 6 and 67 DF,  p-value: 4.276e-05
--------------------------------------------------------------------------------



#rating vs Shelf position
ls=lm(rate~shlf)
summary(ls)
plot(shlf,rate, xlab='Shelf Position', ylab='Rating', main='Rating Vs Shelf Position')
Call:
lm(formula = rate ~ shlf)

Residuals:
    Min      1Q  Median      3Q     Max
-17.157  -9.045  -4.240   5.760  48.175

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  45.8990     3.0092  15.253  < 2e-16 ***
shlf2       -12.4042     4.2021  -2.952  0.00428 **
shlf3        -0.3694     3.7378  -0.099  0.92156
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.12 on 71 degrees of freedom
Multiple R-squared:  0.1503,    Adjusted R-squared:  0.1264
F-statistic: 6.281 on 2 and 71 DF,  p-value: 0.003078
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
```

```
#Sugar content VS Manufacturer

ls=lm(sug~mfr)
> summary(ls)

Call:
lm(formula = sug ~ mfr)

Residuals:
    Min      1Q  Median      3Q     Max
-7.5652 -3.7778  0.0227  3.7087  7.4348

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.000      4.208   0.713    0.478
mfrG           4.955      4.303   1.151    0.254
mfrK           4.565      4.299   1.062    0.292
mfrN          -0.800      4.610  -0.174    0.863
mfrP           5.778      4.436   1.302    0.197
mfrQ           3.143      4.499   0.699    0.487
mfrR           2.857      4.499   0.635    0.528

Residual standard error: 4.209 on 67 degrees of freedom
Multiple R-squared:  0.1445,    Adjusted R-squared:  0.0679
F-statistic: 1.886 on 6 and 67 DF,  p-value: 0.09599

================================================================================
================================================================================
```