Breakfast Cereal Nutrition, Nabisco, and General Mills:

A Statistical Analysis of *Consumer Reports* Breakfast Cereal Ratings and the Nutritional Value of Breakfast Cereal Manufacturers

Stat 445 Project 1

Submitted to: Prof. K. Buhr

Executive Summary

This report uses statistical techniques to answer the following two questions: (i) What qualities of breakfast cereals positively and negatively influence *Consumer Reports* ratings? (ii) The products of which cereal manufacturers received the highest and lowest ratings from *Consumer Reports*. The analysis will be of interest to cereal manufacturers developing new products, consumers who wish to evaluate the meaning of a *Consumer Reports* breakfast cereal rating, and government agencies using *Consumer Reports* ratings to justify regulation of firms within the cereal industry.

Section I of the report addresses the first question. Nine variables are considered as candidates for inclusion in the model. The specification of choice finds strong evidence to support the claims that protein and fiber positively impact the *Consumer Reports* rating of cereals; and that fat, sodium, and sugars negatively impact the rating of cereals. The effect of each variable is found to be economically substantial, statistically significant, and robust across a range of specifications.

Section II of the report addresses the second question. Graphical techniques are first used to identify the likely 'best rated' manufacturer (*Nabisco)* and 'worst rated' manufacturer *(General Mills)*. Next, a model is run to explain ratings using two dummy variables: the first is equal to one for *Nabisco* cereals and the second equal to one for *General Mills* cereals. The effect of each dummy variable is found to be both economically substantial and statistically significant.

1. What breakfast cereal qualities influence Consumer Reports ratings?

1.1: *Background on Consumer Reports*

The *Consumer Reports* mission statement states the following:

> ***Consumer Reports®*** and **ConsumerReports.org®** are published by Consumers Union, an expert, independent nonprofit organization whose mission is to work for a fair, just, and safe marketplace for all consumers and to empower consumers to protect themselves.[1]

Consumer Reports ratings will thus reflect what consumers are seeking in a cereal and what cereal manufacturers may be tempted to misinform consumers about – the cereal's nutritional content.

1.2: *Available Data*

The data consist of the following nutrition variables that may influence *Consumer Reports* ratings:

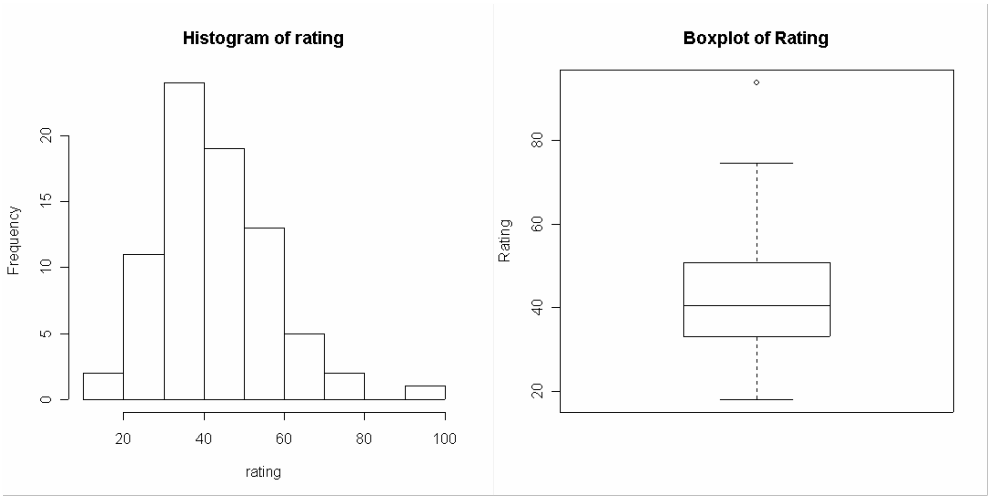
Variable name	***Definition***	***Mean***	***Sd***	***Min***	***Max***
rating	*Consumer Reports* rating	42.67	14.05	18.04	93.70
calories	calories per serving	106.90	19.48	50	160
protein	protein (grams)	2.55	1.09	1	6
fat	fat(grams)	1.01	1.01	0	5
sodium	sodium (milligrams)	159.70	83.83	0	320
carbo	complex carbohydrates (grams)	14.60	3.91	5	23
fiber	dietary fiber (grams)	2.15	2.38	0	14
sugars	sugars (grams)	7.03	4.38	0	15
potass	potassium (milligrams)	98.67	70.41	15	330
vitamins	percentage of FDA recommended daily amount of vitamins and minerals.	28.25	22.34	0	100

* There was one missing value for these variables. ** There were two missing values for this variable.

[1] http://www.consumerreports.org/main/content/aboutus.jsp?FOLDER%3C%3Efolder_id=456983&bmUID=1108875746619

1.3: Outliers in the Rating variable

Before proceeding, the variable **rating** is analyzed to determine whether or not there are any outliers. Thus both a histogram and boxplot of the variable are presented.



Both plots reveal the presence of an outlier on the outer edge of the data. Indeed, eyeballing the data reveals that the highest rating is 93.7 or fully 25.6% higher than the next highest rating of 74.6. To ensure that this outlier does not skew the subsequent analysis, it is excluded from the analysis from this point onwards. However, because it may provide valuable information about what determines **rating**, it will be included in some robustness checks in section 1.8.

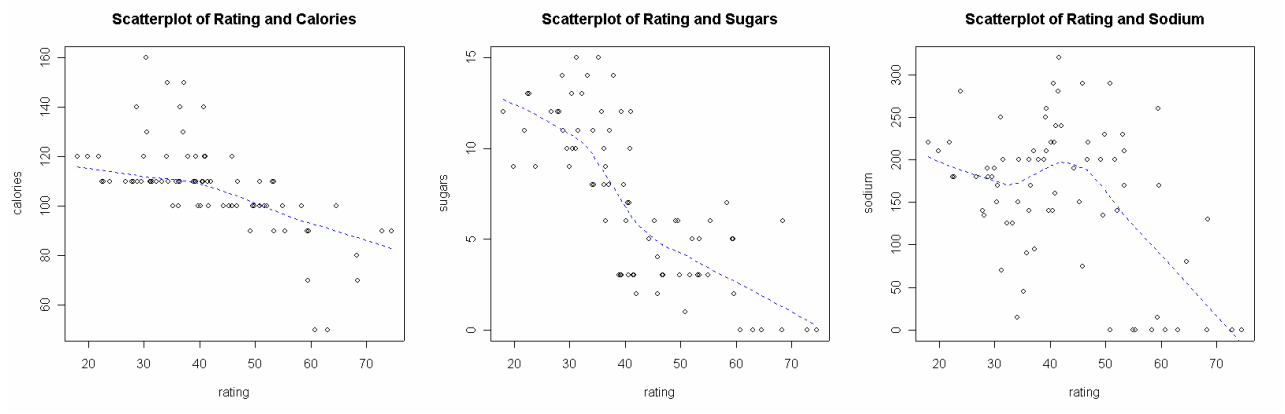
1.4 The Relationship between the candidate independent variables and rating:

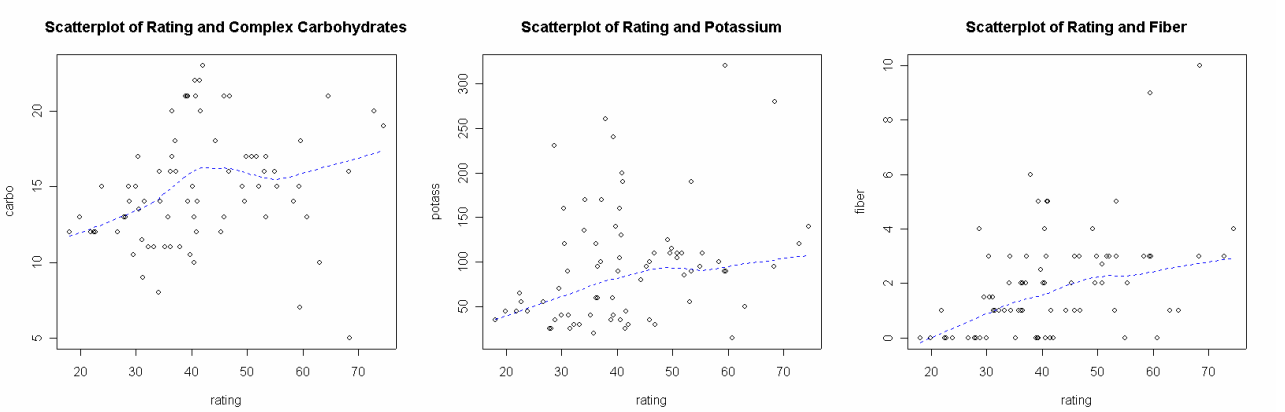
The chart below reveals that most of the candidate independent variables exhibit a substantial correlation with **rating**. None of the variables show a correlation of less than 0.2 and only **vitamins** and **carbo** have correlations less than an absolute value of 0.4, indicating that most of the candidate variables likely impact **rating.**

	calories	**protein**	**fat**	**sodium**	**fiber**
rating	-0.641	0.453	-0.400	-0.430	0.462
	carbo	**sugars**	**potass**	**vitamins**	**rating**
rating	0.206	-0.769	0.301	-0.257	1.000

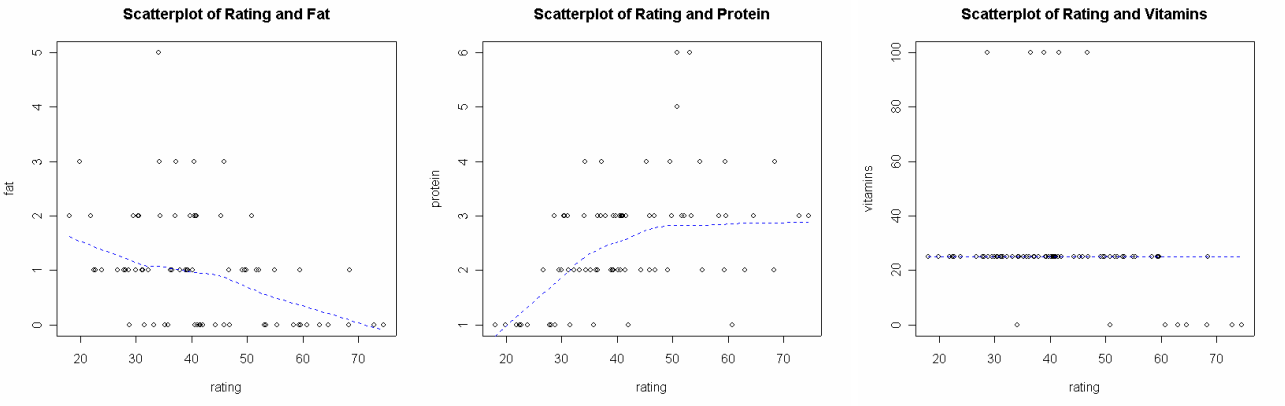
*All variables had similar correlations when the outlier was included in the correlation analysis. The only substantial changes occurred with **carbo** (cor = 0.089), **potass** (cor = 0.416), and **fiber** (cor = 0.584).

To investigate the relationship further, scatterplots are presented with each of the six continuous candidate dependent variables on the y axis and **rating** on the x-axis. The smoothed fitted line through each scatterplot shows the general relationship between the candidate variable and **rating.**





Similar scatterplots are presented for the three categorical variables (fat, protein, and vitamins):



The above graphs demonstrate that there are likely relationships between all of the candidate independent variables and **rating**, with the exception of **vitamins**, which has very little variation as exhibited by the concentration of observations around the 25 mark.

While the above analysis demonstrates that the above candidate independent variables (**vitamins** excluded) do belong in a model where **rating** is the dependent variable, it does not provide an indication of whether there is multicollinearity between the variables.

1.5 The Multicollinearity Problem

In investigating the potential presence of multicollinearity, **calories,** is first considered. Since by definition, calories are functions of the quantities of **protein, fat, carbos,** and **sugars** in a cereal, this variable is unlikely to add any extra explanatory power to a model explaining **rating** that already includes the aforementioned four variables. This suspicion is confirmed when a model is run where **calories** is the dependent variable and **protein, fat, carbos,** and **sugars** are the independent variables. The results of the model are provided below:

Dependent Variable: Calories				
	Estimated Coefficient	**Std. Error**	**t-value**	**p-value**
(Intercept)	-0.181	4.148	-0.04	0.965
sugars	3.950	0.172	23.00	< 0.001
fat	8.615	0.648	13.29	< 0.001
carbo	4.078	0.184	22.15	< 0.001
protein	4.144	0.607	6.83	< 0.001
Multiple R-Squared: 0.9308, Adjusted R-squared: 0.9269				
F-statistic: 235.4 on 4 and 70 DF, p-value: < 0.001				

Note: The p-values provided are those relevant for a two-tailed test.

The key point to note here is the R-squared of 0.9308. This figure means that fully 93.08% of the variation in **calories** is explained by **protein, fat, carbos,** and **sugars**. Additionally all four variables demonstrate an economically substantial and statistically significant (at the 0 level) effect on **calories**. All of this suggests that in a model that includes the four independent variables, **calories**, should be excluded.[2]

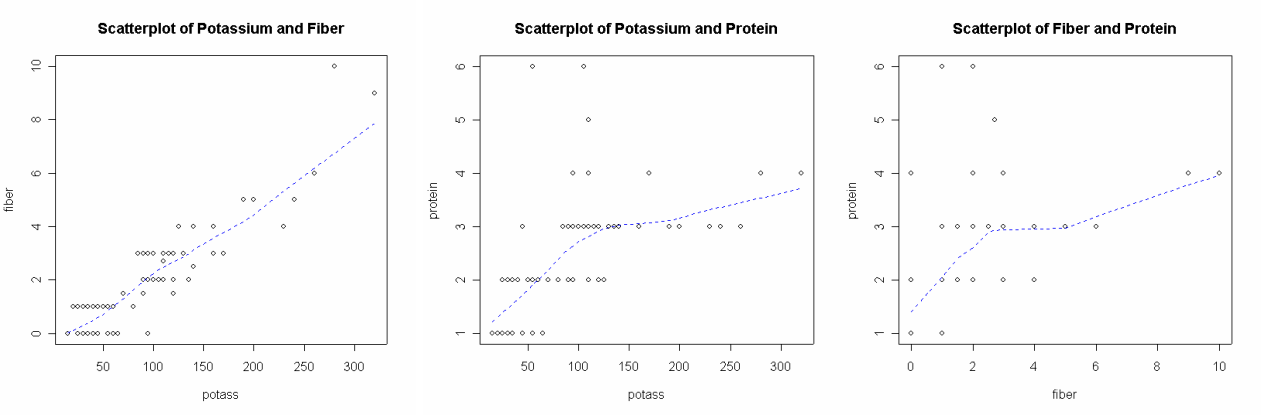
To determine whether there is further multicollinearity, a correlation matrix of the remaining seven candidate independent variables is presented:

	protein	**fat**	**sodium**	**fiber**	**carbo**	**sugars**	**potass**
protein	1.000	0.231	-0.051	0.510	-0.106	-0.270	0.553
fat	0.231	1.000	-0.009	0.103	-0.347	0.288	0.268
sodium	-0.051	-0.009	1.000	-0.067	0.357	0.054	-0.035
fiber	0.510	0.103	-0.067	1.000	-0.315	-0.039	0.914
carbo	-0.106	-0.347	0.357	-0.315	1.000	-0.529	-0.297
sugars	-0.270	0.288	0.054	-0.039	-0.529	1.000	0.083
potass	0.553	0.268	-0.035	0.914	-0.297	0.083	1.000

The following variables appear to exhibit some multicollinearity: **fiber, potass,** and **protein**. Of particular interest is the high correlation of 0.914 between **fiber** and **potass.**

Scatterplots of these variables relating them to one another are presented below and provide further evidence that there is strong collinearity between **potass** and **fiber**:

[2] An alternative would be to include **calories** in place of the other four variables, but this would result in a loss of information on what is driving **ratings.**



1.6.1 Modeling:

Keeping the potentially unaddressed multicollinearity problems in mind, the following model is run:

$$rating = \beta_0 + \beta_1 protein + \beta_2 fat + \beta_3 sodium + \beta_4 fiber + \beta_5 carbo + \beta_6 sugars + \beta_7 potass$$

The model yields the following results:

Dependent Variable: Rating				
	Estimated Coefficient	**Std. Error**	**t-value**	**p-value**
(Intercept)	56.851	1.388	40.95	< 0.001
protein	2.226	0.23	9.66	< 0.001
fat	-3.705	0.218	-16.98	< 0.001
sodium	-0.057	0.002	-23.56	< 0.001
fiber	3.042	0.272	11.19	< 0.001
carbo	0.051	0.07	0.72	0.472
sugars	-1.711	0.06	-28.47	< 0.001
potass	-0.023	0.008	-2.78	0.004
Multiple R-Squared: 0.9867, Adjusted R-squared: 0.9852				
F-statistic: 686.7 on 7 and 65 DF, p-value: <0.001				

Note: The p-values provided are those relevant for a one-tailed test.

Dropping **potass** (because it is extremely related to **fiber** and has a very small coefficient of the opposite than expected sign) reveals results that are similar for all of the above variables. Of special interest, however, is the variable **carbo**, which has almost no statistical significance (p-value of 0.472 and t-value of 0.07) in the resulting regression. Thus, for the final specification, which is the specification of choice, this variable is dropped.[3]

[3] It is noted that **carbo** was used in the regression justifying the dropping of calories. The specification is also run where **carbo** is dropped, but **calories** is included. The regression showed an economically small (-0.02) and statistically insignificant value (p-value: 0.07) on the **calories** coefficient, indicating that even when **carbo** is excluded, **calories** does not help explain **ratings**.

1.6.2 The Specification of Choice:

The specification of choice includes **rating** as the dependent variable and **protein**, **fat, sodium, fiber,** and **sugars** as the independent variables. It is represented by the following equation:

$$rating = \beta_0 + \beta_1 protein + \beta_2 fat + \beta_3 sodium + \beta_4 fiber + \beta_5 sugars$$

The results of this regression are presented below:

Dependent Variable: Rating				
	Estimated Coefficient	**Std. Error**	**t-value**	**p-value**
(Intercept)	57.833	0.752	76.89	< 0.001
protein	2.045	0.229	8.94	< 0.001
fat	-3.918	0.208	-18.85	< 0.001
sodium	-0.057	0.002	-24.7	< 0.001
fiber	2.359	0.113	20.88	< 0.001
sugars	-1.778	0.049	-36.01	< 0.001
Multiple R-Squared: 0.9851, Adjusted R-squared: 0.984				
F-statistic: 884.4 on 5 and 67 DF, p-value: <0.001				

Note: The p-values provided are those relevant for a one-tailed test.

This reveals that each cereals rating can be approximated by the following formula:

$$rating = 57.833 + 2.045\,protein - 3.918\,fat - 0.057\,sodium + 2.359\,fiber - 1.778\,sugars$$

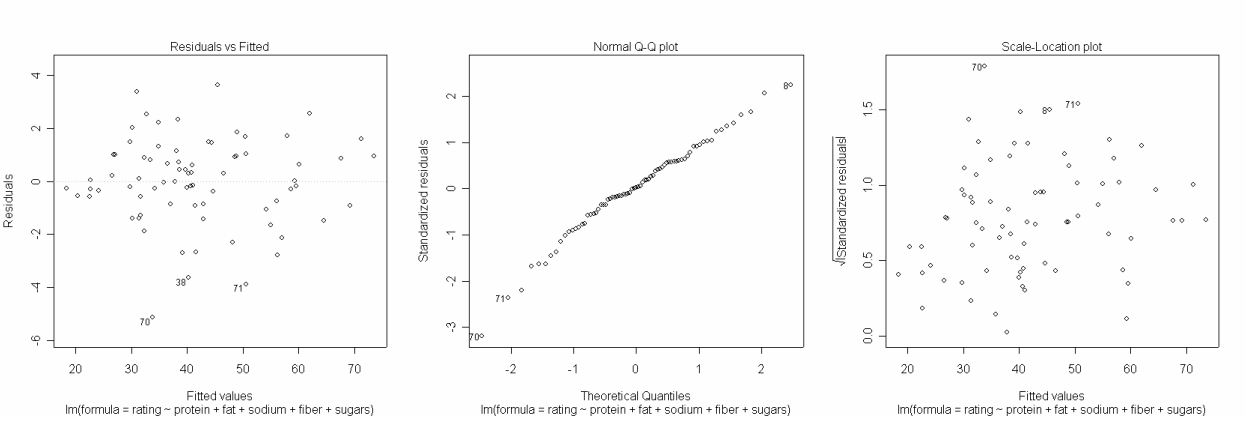
This regression demonstrates some powerful results. The five independent variables chosen explain 98.51% of the variation in **ratings**. Furthermore, all the p-values indicate substantial evidence of the following: (i) **protein** and **fiber** have a positive and statistically significant effect on **ratings**; and (ii) **fat**, **sugars**, and **sodium** have a negative and statistically significant effect on **ratings**.

With 95% confidence the following five claims can be made about the effect of each independent variable on **rating:** (i) A 1 gram increase in the amount of protein in a cereal serving will increase the rating by an amount between 1.587 and 2.503; (ii) A 1 gram increase in the amount of fiber in a cereal serving will increase the rating by an amount between 2.133 and 2.585; iii) A 1 gram increase in the amount of fat in a cereal serving will decrease the rating by an amount between 3.502 and 4.334; iv) A 1 gram increase in the amount of sodium in a cereal serving will decrease the rating by an amount between 0.053 and 0.061; (v) A 1 gram increase in the amount of sugars in a cereal serving will decrease the rating by an amount between 1.680 and 1.876.

1.7 An Analysis of the Residuals

The diagnostic plots below reveal that the residuals behave very well. The Residuals vs. Fitted values plot and the Scale-Location plot indicate that the size of the residuals remain constant across all magnitudes of the fitted values and thus exhibit no systematic relationship to the fitted

values. The Normal QQ Plot reveals that the residuals are approximately normally distributed. Of slight concern may be the presence of the two outliers, labelled as points 70 and 71. These points stray quite far from what would be an otherwise straight line. However, their presence is not overly worrisome because the sample is of large enough size that we may expect two outliers to be present.



1.8 Robustness Checks

Running the above model with the outlier included in the analysis reveals no substantial change in any of the coefficients on any variable and does not impact the R-squared values. Additionally, the results of a model that includes all nine candidate independent variables (i.e. **protein**, **fat**, **sodium**, **fiber**, **carbo**, **sugars**, **potass**, **vitamins**, and **calories**) does not differ substantially from the specification of choice, although the adjusted R-squared does increase to 1. The coefficients on all variables maintain their sign. The only significant difference is that the coefficients on **fat** and **sugars** declines substantially, which is due to their multicollinearity with **calories**.

Question 2: The products of which cereal manufacturer received the highest and lowest ratings from Consumer Reports?

2.1 Representation of Manufacturers

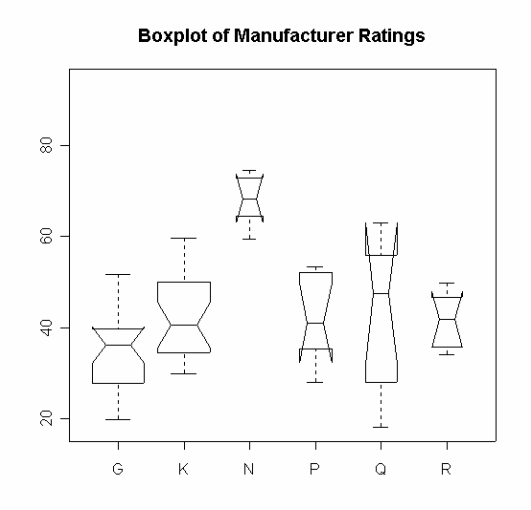
For the second question of this report, the differential mean ratings of cereal manufacturers are analyzed to determine if there are companies that produce cereals with significantly higher or lower ratings than the others.

The cereal manufacturers are represented as follows:

G = *General Mills*	K = *Kellogs*	N = *Nabisco*
P = *Post*	Q = *Quaker Oats*	R = *Ralston Purina*

American Home Food Products is not considered. It only produces one cereal that was tested.

The graph below is a boxplot of each of the included manufacturers. The box widths are proportional to the number of cereals each manufacturer produces. The notches are rough indicators of a 90% confidence interval about the mean product rating for each manufacturer.



The plot reveals that two firms that may be significantly different from the rest: *Nabisco* appears to have received substantially higher ratings than the other manufacturers, while *General Mills* appears to have possibly received lower ratings than the other manufacturers.

To test the hypothesis, a model is run, where dummy variables are included for both *Nabisco* and *General Mills* cereals. The model is as follows:

$$Rating = \beta_0 + B_1 GenMills + B_s Nabisco$$

The results of the model are presented below:

Dependent Variable: Rating				
	Estimated Coefficient	**Std. Error**	**t-value**	**p-value**
(Intercept)	43.24	1.633	26.482	< 0.001
GenMills	-8.754	2.933	-2.984	0.002
Nabisco	24.729	4.944	5.002	< 0.001
Multiple R-Squared: 0.3544, Adjusted R-squared: 0.338				
F-statistic: 20.4 on 2 and 74 DF, p-value: <0.001				

Note: The p-values provided are those relevant for a one-tailed test.

The p-values indicate that there is ample evidence to support the claims that (i) **General Mills** cereals receive on average statistically significantly lower ratings than the cereals of the other five companies considered; and (ii) **Nabisco** cereals receive on average statistically significantly higher ratings than the cereals of the other five companies considered. The point estimates reveal that **General Mills** cereals receive an average rating 8.754 points less than other cereals and **Nabisco** cereals receives an average rating 24.729 points higher than other cereals. With 95% confidence the following two claims can be made: (i) **General Mills** cereals receive an average rating between 2.888 and 14.620 points less than the ratings of other cereals; and (ii) **Nabisco** cereals receive an average rating between 14.841 and 34.617 points higher than the ratings of the other cereals.