□□□□□□□□□□□□□□□□□□□□□□□□□□□□

STAT445 Midterm Project1

□□□□□□□□□□□□□□□□□□□□□□□□□□□□

Executive Summary

This report works on the dataset of "Part of This Nutritious Breakfast!" In this dataset, 77 different breakfast cereals were collected. The dataset also explores the nutrition components, type of serving and providers' information, which enables us to do analysis on a variety of problems.

Firstly, we will look at different variables separately to get a rough idea of the variable. Descriptive statistics and graphical plots or boxplots are very helpful and heavily used for this purpose. We have found the different market shares of providers, serving type, serving size, descriptive statistics of the nutrition contents and consumer ratings. It is also interesting to know rating fits a gamma distribution well. Secondly, we want to find how variables interact with each other. Obviously, we can try to test a lot of relationships between these 16 variables but we are going to focus on two main relationships in this report. The first question is what determines calories? The second is what determines rating? Therefore, we will do some exploratory analysis on these two types of relationships. Thirdly, we need regression techniques to show to what extend these variables are interacted and whether the interaction is significant or not. We have found from the regressions that protein, fat, sugars and carbohydrates have significantly positive effects on the calories. This result is also quite intuitive because we've already known that the calories, a measure of energy mainly come from nutrition contents such as protein, fat, sugars and carbohydrates. We also have an interesting finding that the effects of the nutrition contents are bigger if they work together. We further look at how the rating interacts with other variables. One caveat here is to include calories and protein, fat, sugars or carbohydrates in the single regression model because calories are already shown to be determined by protein, fat, sugars or carbohydrates. Therefore, we can choose whether to drop calories or to drop those concerned nutrition contents in a single regression model. It can be seen that the model with protein, fat, sugars and carbohydrates but without calories works better than the model with calories but without the four nutrition contents in terms of significance levels and model fitness.

The report is organized as follows: Part I provide a brief description of the dataset; Part II explores the relationship between variables; Part III gives regression results; Part IV concludes the report.

Part I Description of the Dataset

The dataset collected 77 different cereals. However, there are 4 missing values in the dataset, one for sugar, one for carbohydrates and two for potassium. It is reasonable for us to omit the observations with missing values and we end up with 74 observations.[1] The following descriptive statistics are based on the new dataset with 74 observations. They are provided by 7 different manufacturers, among which *Kelloggs* has the biggest market share of 31%, followed by *General Mills* with 30%, *Post* with 12%, *Quaker Oats* and *Ralston Purina* each for 9%, *Nabisco* with 7% and *American Home Food Products* with 1%. The description of the characteristics of the cereals is summarized in Table 1.

[1] The 58th observation has a missing value for sugars and carbohydrates. The 5th observation has a missing value for potassium and the 21st observation also has a missing value for potassium. Therefore, these 3 observations are omitted

Table 1 Characteristics of the Cereals (per serving)

Nutrition Contents	Mean	Median	Standard Deviation
Protein (g)	2.51	2.50	1.08
Fat (g)	1.00	1.00	1.01
Carbohydrates (g)	14.73	14.50	3.89
Sugars (g)	7.12	7.00	4.36
Sodium (g)	162.40	180.00	82.77
Fiber (g)	2.18	2.00	2.42
Potassium (mg)	98.51	90.00	70.88
Vitamins	29.05	25.00	22.29
Calories	107.00	110.00	19.84
Other Variables		Ratio Statistics	
Type of serve:			
-cold		99%	
-hot		1%	
Shelf in the market			
-1		26%	
-2		27%	
-3		47%	
Weigh			
-less than 1 ounce		4%	
-1 ounce		82%	
-more than 1 ounce but less than 1.5 ounces		14%	
Cups			
-less than 1 cup		55%	
-1 cup		39%	
-more than 1 cup but less than 1.5 cups		5%	
Rating	Mean	Median	Standard Deviation
Consumer Rating	42.37	40.25	14.03

The above table gives us a rough idea of different variables in the dataset. After ploting, boxploting and drawing histograms of the variables in the dataset[2], I find that “rating” can fit a gamma distribution with shape equal to 9.79 and rate equal to 0.23. However, other variables do not show any obvious fit of a particular distribution.

Part II Explore the Relationship between Variables and Research Questions

The dataset enables us to test two main types of relationship between variables. The first relationship is between calories and nutrition contents such as protein and sugars. The second relationship is between the consumer rating and the characteristics of the cereals.

It is well-known that calories, a measure of energy come from certain nutrition contents such as protein, fat, carbohydrates and sugars. Some microelements like fiber, sodium and potassium do not directly contribute to the calories. After knowing this we can have an interesting question in mind: at what extent certain nutrition contents affect energy? At this part of the report, we will

[2] The plots, boxplots and histograms of the variables in the dataset are not shown here.

explore the relationship between calories and contributative nutrition contents one by one with the help of a series of plots in Figure 1.[3]

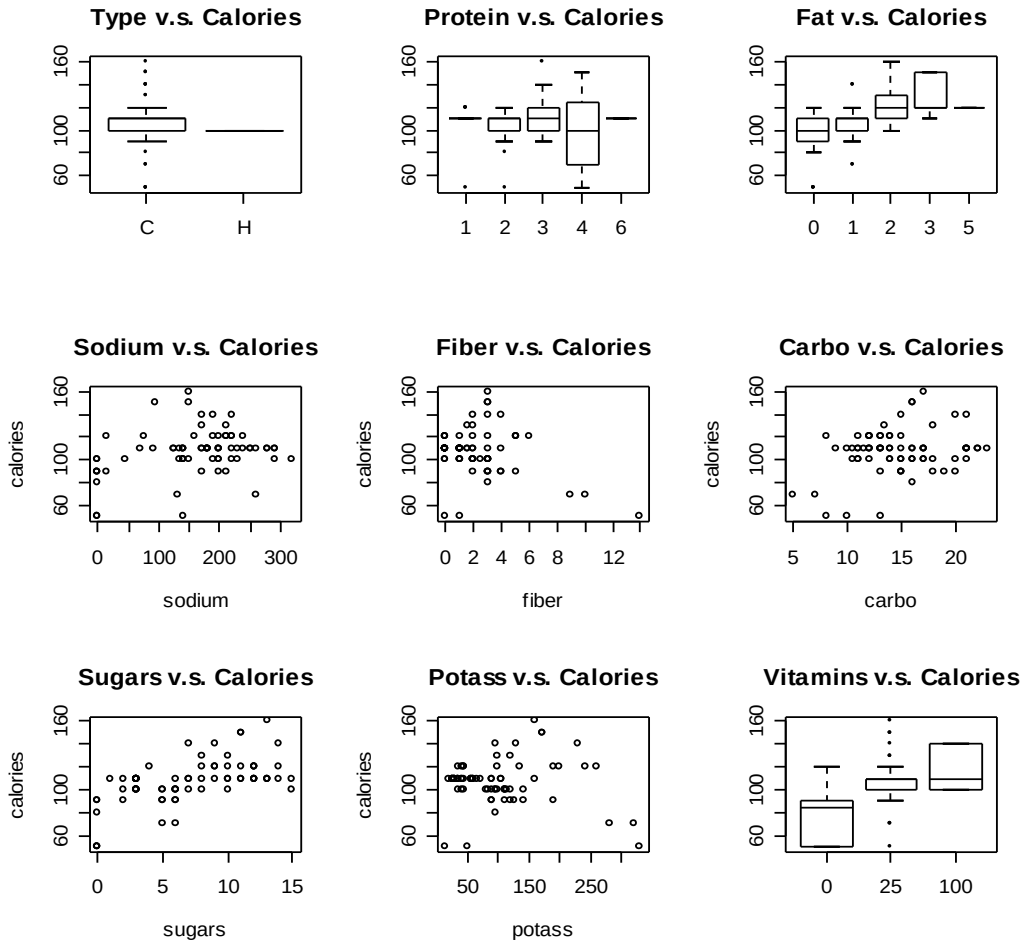

Figure 1

It can be seen from Figure 1 that nutritious contents like protein and fat show a clear pattern of movement with calories. Some microelements such as potassium and sodium do not give us a clear relationship with calories. To decide the relationship between these variables, we still need to use regression analysis in Part III.

The consumer's rating matters for both manufacturers and consumers. Manufacturers might adjust their production plans to the consumer's rating while consumers might take the consumer's rating as a guideline when making their own purchasing decisions. Therefore, it is important to know what factors actually have an effect on consumers' rating report. Figure 2 shows the relationship between rating and other variables in the dataset.

Some variables such as protein, fat, sugars fiber have demonstrated positive or negative relationship to rating as depicted in various plots or boxplots in Figure 2. We will explore more about the relationship in our regression analysis. There are also other types of relationships between variables, e.g. "weigh" has a significant positive effect on "calories", but they are not

[3] I use "boxplot" when the variables on the axis have few discrete values such as "type", "protein", "fat" and "vitamins". I use "plot" when the variables on the axis have a wide range of discrete values such as "sodium", "fiber", "carbo", "sugars" and "potass".

highly relevant to our research questions. Therefore, only the plots helpful to answer the research questions are shown in the above figures.

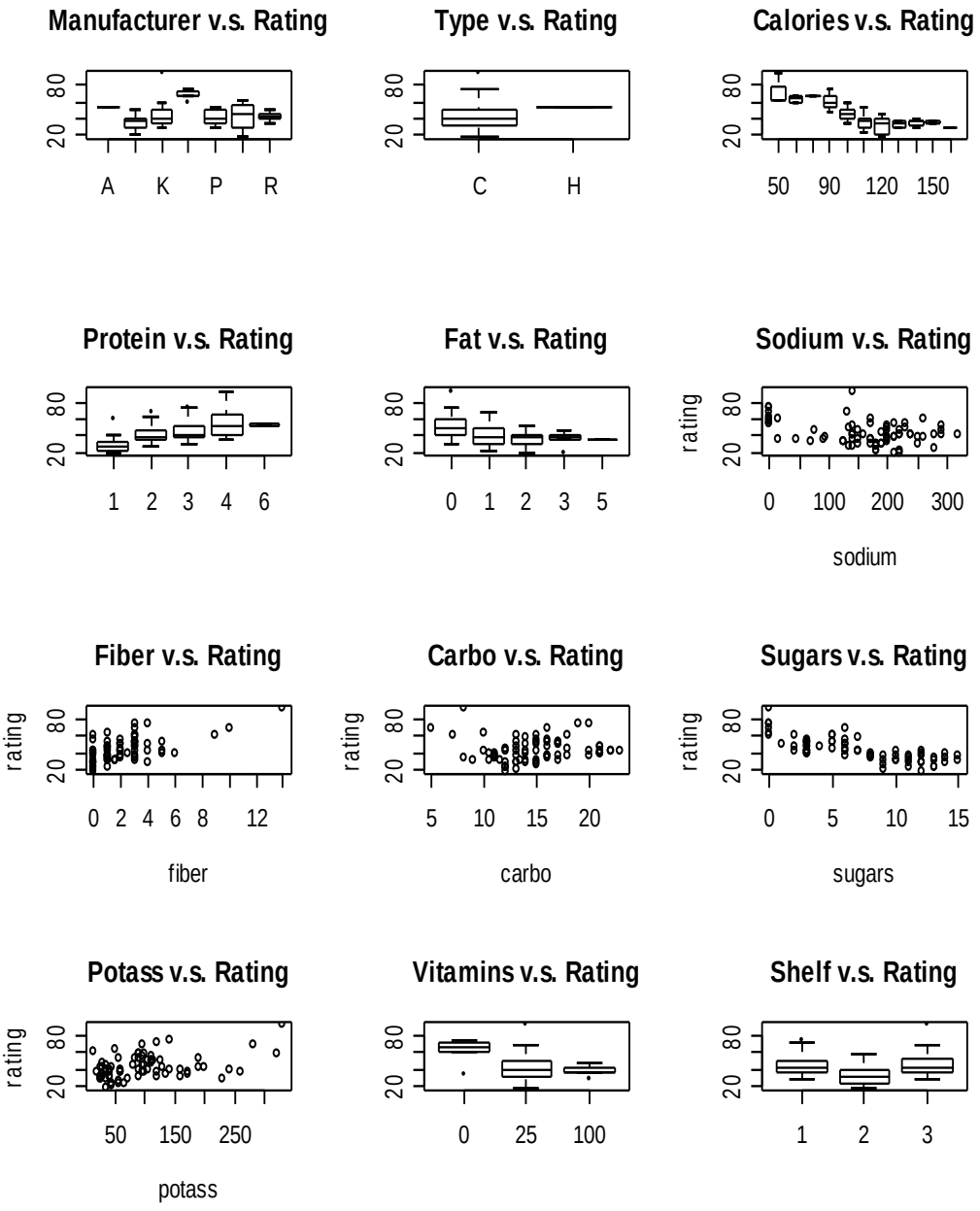

Figure 2

Part III Regression Statistics

In this part, we will use regression techniques to answer the two research questions in the form of two regression tables.

Table 2 Regression of Calories

Dependent Variable: Calories					
	(1)	(2)	(3)	(4)	(5)
Intercept	97.03***	105.45***	88.61***	86.70***	
	(2.83)	(5.93)	(3.67)	(8.81)	
Fat	10.00***				8.68***
	(2)				(0.65)
Protein		0.63			4.10***
		(2.17)			(0.61)
Sugar			2.59***		3.94***
			(0.44)		(0.17)
Carbohydrates				1.38*	4.06***
				(0.58)	(0.17)
Adjusted R-square	0.25	-0.01	0.31	0.06	0.94

*Note: Standard deviation is reported in the bracket.

From Table 2, we can see to what extent the nutrition contents affect the energy level. Regressions (1) to (4) are regression of calories on fat, protein, sugar and carbohydrates separately. The coefficients (except fat, which is slightly bigger) are smaller than the coefficient of the fifth regression where I regress the calories on four nutrition variables in a single linear model. This implies nutrition contents working together can have a bigger effect on calories than those working separately. To check the assumptions of the linear model, plots on the residuals of the fifth regression is shown below in Figure 3.

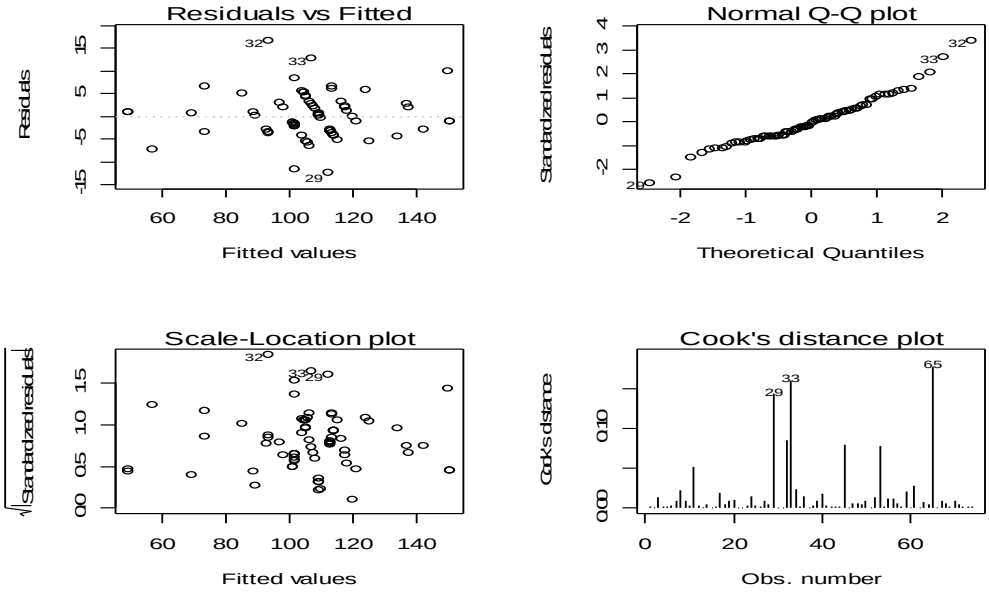

Figure 3

We can find that the assumptions of linear regression are not violated in Figure 3 and thus it is safe to draw inference from the regression results in Table 2.

We will then move on to the relationship between rating and other variables, followed by

some plots to check if the assumptions of linear model are violated.

Table 3 Regression of Rating

Dependent Variable: Rating			
	(1)	(2)	(3)
Intercept	54.93***	78.39***	55.59***
	(0)	(5.93)	(1.03)
Calories	-0.22***	-0.32***	
	(0)	(0.06)	
Protein	3.27***		2.34***
	(0)		(0.17)
Fat	-1.69***		-3.73***
	(0)		(0.16)
Sodium	-0.05***	-0.04**	-0.05***
	(0)	(0.01)	(0)
Fiber	3.44***	3.89**	3.16***
	(0)	(1.14)	(0.17)
Carbohydrates	1.09***		0.15**
	(0)		(0.05)
Sugars	-0.72***		-1.64***
	(0)		(0.04)
Potassium	-0.03***	-0.05	-0.03***
	(0)	(0.04)	(0)
Vitamins	-0.05***	0	-0.05***
	(0)	(0.05)	(0)
Adjusted R-square	1	0.68	0.99

*Notes: Standard deviation is reported in the brackets.

In table 3, three linear models are regressed to test what affect the rating. In the first regression, I regress rating on calories and other nutrition contents. All the coefficients are significant. Protein, fiber and carbohydrates have positive effect on rating while calories, fat, sodium, sugars, potassium and vitamins have negative effect on the rating. The adjusted R-square is as high as 1. However, we have shown that calories can also be determined by the levels of protein, fat, sugars and carbohydrates. If we use both calories and its determinants in the same regression we would incur the problem called “multicollinearity” in the econometrics literature. Therefore, in the second regression, I drop protein, fat, sugars and carbohydrates and use calories to be a representative of them. The performance of the second regression in terms of the significance level and adjusted R-square is not as good as the third regression where I use protein, fat, sugars and carbohydrates as a representative for calories. Therefore, the third regression is preferred here. We will also check the assumptions of the linear model for the third equation in the following figure.

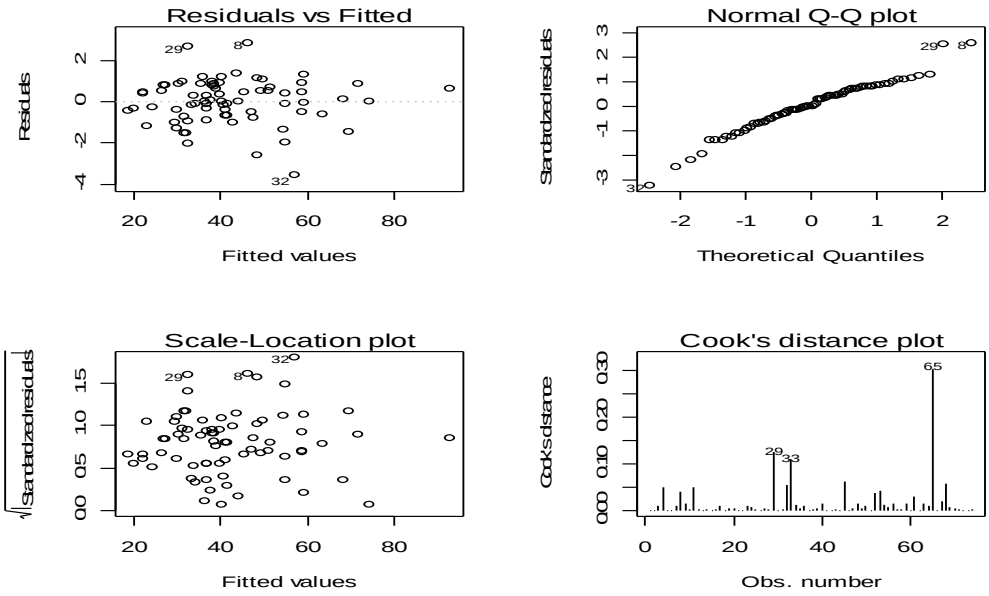

Figure 4

Part IV Conclusions, Limitations and Further Questions

After the first three parts analysis of the data, we can have a clearer view of the dataset. The plots and boxplots in the second part give us the moving pattern between calories with other nutrition variables and between rating and other nutrition variables. The third part of the regressions further confirms the relationships between different variables. We have found that calories are mainly from protein, fat, sugars and carbohydrates, which conform to our knowledge of nutrition. We have also found that rating is positively correlated with protein, fiber and carbohydrates and negatively correlated with calories, fat, sodium, potassium, vitamins and sugars. These findings are good for manufacturers because they can adjust their production to the consumer rating and thus more demand of their products. Consumers can also benefit from the findings because they know what kind of nutrition is valued much in the cereals, from which they can have guidelines of their purchasing choices.

There are some limitations in the dataset, for example, the dataset could be enlarged to include more observations, more nutrition contents and more information on sales of the cereals. Further research questions would be to test whether these nutrition contents have an effect on sales if the dataset allows and to use regression techniques other than linear regression, such as Maximum Likelihood or some nonparametric methods, which impose fewer assumptions on the distribution of the error term.