UNIVERSITY OF NEW SOUTH WALES
SCHOOL OF MATHEMATICS AND STATISTICS

MATH3821 Statistical Modelling and Computing
Term Two 2020

Assignment Two

Given: Friday 17th July 2020 **Due date: Sunday 2nd August 2020**

INSTRUCTIONS: This assignment is to be done **collaboratively** by a group of **5 students**. The same mark will be given for the report to each student within the group, unless I have good reasons to believe that somebody did not do anything.

You will need to produce and submit a report of your work in PDF format. This report will not contain more than 10 pages, excluding the Appendix that should contain your computing codes. The report is due 11:59 pm, Sunday 2nd August. The first page of this PDF should be **this page**. Only one of the five students should submit the PDF file on Moodle, with the names of the other students in the group clearly indicated in the document.

I/We declare that this assessment item is my/our own work, except where acknowledged, and has not been submitted for academic credit elsewhere. I/We acknowledge that the assessor of this item may, for the purpose of assessing this item reproduce this assessment item and provide a copy to another member of the University; and/or communicate a copy of this assessment item to a plagiarism checking service (which may then retain a copy of the assessment item on its database for the purpose of future plagiarism checking). I/We certify that I/We have read and understood the University Rules in respect of Student Academic Misconduct.

Name	Student No	Signature	Date
Jiaxuan Li	z5142142	Jiaxuan Li	02/08/2020
Shourong Lin	z5135064	林守蓉	02/08/2020
Yuting Zhao	z5163413		02/08/2020
Fangyue Chen	z5175111		02/08/2020
Huiyuan Zheng	z5192056		02/08/2020

Introduction & Abstract

Having a bowl of cereals is a good way to start a day since it is relatively nutritious and easy to make. However, choosing the right nutritious breakfast cereals is not always easy. In this study, we are going to investigate 77 commonly available breakfast cereals (based on the information on newly mandated F&DA food label) to explore all the explanatory variables that have relationships with the rating by statistical analysis.

In this study, we analyse the cereals dataset by both model fitting and data visualizations. Moreover, we use the Bootstrap method as our simulation strategy. In the end of this report, we will talk about the limitations and drawbacks of our models. And we will state the facts and interesting findings that we found after data analysis.

Key Assumption

Assumptions will be made throughout the report. Here, a list of the assumptions is provided:

Assumption of Data Collection

1. 77 available breakfast cereal are chosen randomly

2. All records of are accurate and precise during data collection process

Assumption of model

Standard assumptions for linear regressions, which include linearity, homoscedasticity, independence, and normality of errors.

The data collection assumptions are made before any analysis and would be taken for granted. The model assumptions can be examined through diagnostic plots and some preliminary analysis.

Goal

We will figure out what nutritional content scores high in rating and provide customer a criterion to choose a right healthy cereal. Moreover, we will answer three questions that provide useful information:

1. Based on the distribution of recommended calorie intake, are there any valuable relationship between sugar and rating.
2. Are there any breakfast cereals in market are disparate or virtually identical? Do manufactures have different strategies when adding depth of product line?
3. Are most of the cereals in the market are healthy that can be a good choice for those trying to lose weight?

Data cleaning and wrangling

The dataset contains per-serving nutritional information and grocery shelf location for 77 breakfast cereals from seven manufacturers. We have 16 variables in dataset (four categorical and twelve quantitative).

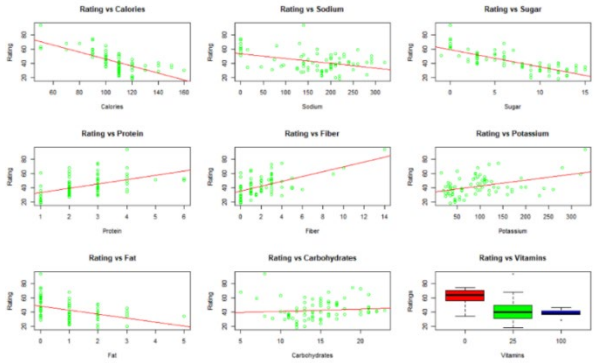
1. We change the type of column "type", "mfr" and "shelf" to factor

2. We remove three unknown values which has -1 as missing value for potassium, complex carbohydrates and sugars. Therefore, Almond Delight, Cream of Wheat (Quick), Quaker Oatmeal are excluded in our analysis, the number of data change from 77 to 74. Note that for the remaining 74 cereals, only one cereal, Maypo, served as Hot. Hence, we need to put special attention to this data point as it may be a possible influential observation.

3. We added a new column mfr_full to store the full manufacturer name to make our graphs clearer when we investigate the manufacturer variable.

Exploratory Data Analysis

1. Correlation interpretations



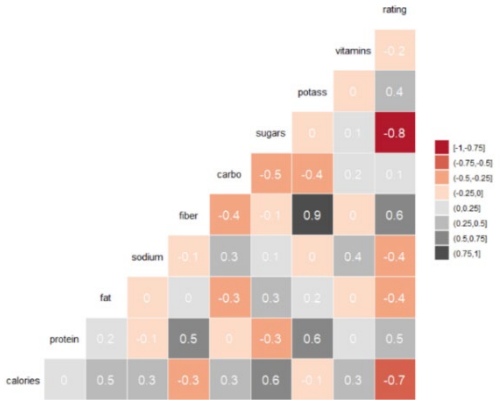


Figure 1 – Plots between Rating and the 9 Exploratory variables
Figure 2 – Correlation matrix of Rating and 9 Exploratory variables

We first examine the relationship between the response rating and the exploratory variables (nutritional content), which includes calories, protein, fat, sodium, potass, carbo, fibre, sugars and vitamins. To visualize the associations, we drawn the scatterplot for each exploratory variable vs rating. Note that vitamins only have three discrete values 0, 25 and 100. Hence it is better to draw a boxplot for vitamins vs rating. Check Figure 1 above.

From Figure 1, we can easily see the trends. It is obvious that protein, fiber and potassium have a positive correlation with rating, which means that rating increases as these variables increase. This also makes sense since by common sense since we know that fiber, protein and potassium are the good nutrients for our body. Similarly, we can see that calories, sodium, sugar, fat and vitamins have a negative correlation with rating.

From Figure 2, it shows the exact correlation coefficients of rating and exploratory variables. By reading the table, we notice that sugars, fat, sodium and calories are negatively correlated to rating, with coefficients -0.8, -0.4, -0.4 and -0.7 respectively. On the other hand, potassium, fiber and protein are positively correlated to rating, with coefficients 0.4, 0.6 and 0.5 respectively. Hence we get the idea that cereals with high fiber, high protein and high potassium may get a high rating by this grading system and vice versa for those have negative correlation with rating.

For instance, All-Bran with Extra Fiber, which has highest rating of 93.70, has the highest fiber and high protein content without fat and sugars. On the other hand, Cap'n'Crunch, which has the lowest rating of 18.04, has high sugars and high fat, and low in protein and no dietary fiber at all. These are the two extreme rating values that shows how nutritional contents affect the score on rating.

Moreover, notice that fiber and potassium have a positive 0.9 correlation. We get the intuition that these two predictors might not be independent, so we need to put special attention on them when fitting a model.

2. Model fitting

```
> fit_full <- lm(rating~mfr+type+calories+protein+fat+sodium+fiber+sugars+
+                potass+vitamins+shelf+weight+cups, data = cereals)
> summary(fit_full)

Call:
lm(formula = rating ~ mfr + type + calories + protein + fat +
    sodium + fiber + sugars + potass + vitamins + shelf + weight +
    cups, data = cereals)

Residuals:
     Min       1Q   Median       3Q      Max
-2.42382 -0.69940 -0.00571  0.70084  2.06068

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 55.076727   1.825597  30.169  < 2e-16 ***
mfrG         1.442114   1.284998   1.122 0.266537
mfrK         1.648760   1.328530   1.241 0.219765
mfrN         1.642117   1.357221   1.210 0.231396
mfrP         1.191904   1.344466   0.887 0.379126
mfrQ         0.971365   1.309519   0.742 0.461325
mfrR         3.194190   1.382373   2.311 0.024559 *
typeH             NA         NA      NA       NA
calories    -0.077273   0.021299  -3.628 0.000619 ***
protein      2.549033   0.180105  14.153  < 2e-16 ***
fat         -3.275173   0.234439 -13.970  < 2e-16 ***
sodium      -0.054029   0.002297 -23.518  < 2e-16 ***
fiber        2.823395   0.195091  14.472  < 2e-16 ***
sugars      -1.658568   0.047374 -35.010  < 2e-16 ***
potass      -0.028189   0.006691  -4.213 9.25e-05 ***
vitamins    -0.047219   0.007176  -6.580 1.71e-08 ***
shelf        0.069290   0.220138   0.315 0.754117
weight       7.816710   2.182893   3.581 0.000717 ***
cups         1.459306   0.687006   2.124 0.038087 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.078 on 56 degrees of freedom
Multiple R-squared:  0.9955,    Adjusted R-squared:  0.9941
F-statistic: 724.3 on 17 and 56 DF,  p-value: < 2.2e-16
```

```
> AIC.model_summary<-summary(AIC.model)
> AIC.model_summary

Call:
lm(formula = rating ~ sugars + fiber + sodium + fat + protein +
    vitamins + potass + carbo + calories, data = cereals)

Residuals:
       Min         1Q     Median         3Q        Max
-5.343e-07 -2.537e-07  3.961e-08  2.424e-07  5.513e-07

Coefficients:
             Estimate Std. Error    t value Pr(>|t|)
(Intercept)  5.493e+01  2.794e-07  196559702   <2e-16 ***
sugars      -7.249e-01  3.311e-08  -21895192   <2e-16 ***
fiber        3.443e+00  4.756e-08   72399805   <2e-16 ***
sodium      -5.449e-02  4.910e-10 -110974232   <2e-16 ***
fat         -1.691e+00  8.101e-08  -20877762   <2e-16 ***
protein      3.273e+00  5.551e-08   58964906   <2e-16 ***
vitamins    -5.121e-02  1.779e-09  -28778552   <2e-16 ***
potass      -3.399e-02  1.601e-09  -21228850   <2e-16 ***
carbo        1.092e+00  3.492e-08   31287364   <2e-16 ***
calories    -2.227e-01  7.501e-09  -29694282   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.069e-07 on 64 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 1.696e+16 on 9 and 64 DF,  p-value: < 2.2e-16
```

We are fitting a multiple linear regression of rating as the response and all other variables as predictors. By looking at the summary output for this full model, we found mfr, type, and shelf are not significant at 5% level of significance. Hence, we drop these predictors.

We fit a new model with the remaining predictors, which include mfrR, calories, protein, fat, sodium, sugars, potass, vitamins, weigh and cups. And we fit a model with only intercept (rating~1) as well. Then we use these two models to run the stepwise AIC forward selection as our model selection strategy. From the R output, we obtain the final model "rating ~ sugars + fiber + sodium + fat + protein + vitamins + potass + carbo + calories" as it has the smallest AIC value. Hence, the equation for our final model is:

rating = 54.93 - 0.22*calories + 3.27*protein - 1.69*fat - 0.054*sodium + 3.44*fiber + 1.09*carbo -0.72*sugars - 0.034*potass - 0.051*vitamins

We can extract information by reading the coefficient of each variable, for example, one-unit change in fiber leads to 3.44-unit multiplicate change in rating. The result of the final model is consistent with the correlation matrix in Figure2 and the scatterplots in Figure 1. From the graphic and the final model, we learn that the rating is highly and positively correlated with fiber and protein, while it is highly and negatively correlated with fat and sugars. The remaining variables have weak relationships with rating. Now we get the idea that cereals with high ratings is associated with high fiber, high protein, low fat, and low sugars.

Our final model gives a perfect fit with R-squared strictly equals to 1. The p-values of t-tests on each predictor are 2.2e-16, which means that they are statistically significant at 5% level. As R-squared = 1 indicates that 100% of variability is explained by this model.

3. Diagnostics

For an appropriate linear model, it needs to satisfy a few assumptions, which are

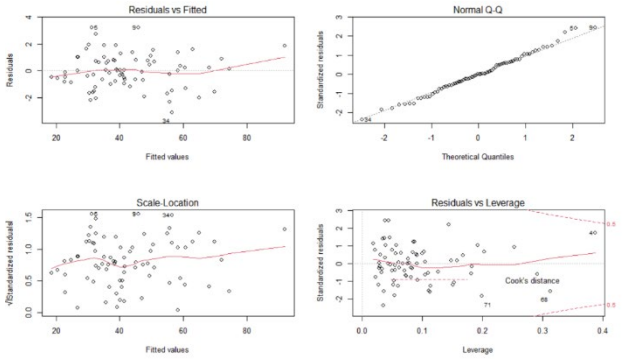
- Linearity: The relationship between predictors and the mean of response is linear.
- Homoscedasticity: All errors have the same variance.
- Independence: Observations are independent of each other.
- Normality of errors.

To detect the violations of assumptions, we examine the diagnostic plots.

In the residual plot, we do not see the U shape, so linear assumption doesn't seem to be violated. There is no fan shape as well, which means our variance is constant for different fitted values, so it also meets the homoscedasticity assumption. The normal quantile plot shows no evidence that the normality assumption is violated as the linear line captures most of the data points. Hence, no transformations were needed to fix a violation of assumptions.

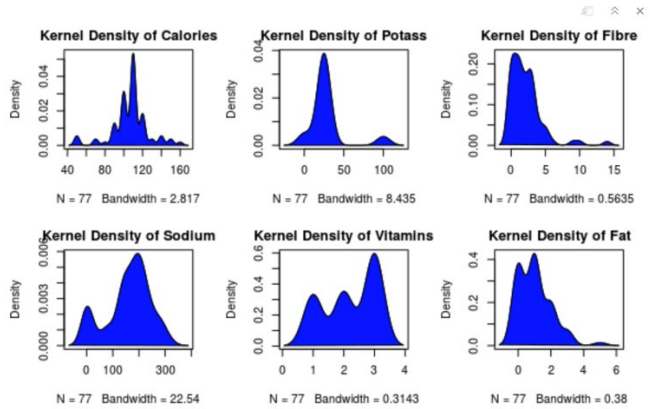
From the leverage plot, it is obvious that there is no outlier as all data points are not outside the cook's distance red lines. Moreover, we do not see any data point with high leverage and high residual, which means that there is no influential point as well. No data points need further investigation.

Therefore, by examining all diagnostic plots, we can conclude that this multiple linear regression is an appropriate model as it does not violate the assumptions.



4. Kernel Density

Kernel density plots present the distribution of data for a period or continuous time interval., and the bandwidth performs as a smoothing parameter to smooth out the noise, determining the trade-off between bias and variance. If the bandwidth is large, the density tends to be smoother and its bias is high, for example, in the kernel density of Potassium and Sodium, a small peak on the plot of Sodium exists at approximately 0 and a large peak at around 200 , the bandwidth is 22.54, which results in high-bias and underfitting probably, when the bandwidth is larger, the amount peaks on the plot is less; if the value of bandwidth is low, as the kernel density of calories and fibre shown, there are more fluctuations and distributed widely along the axis, these two plots include more peaks, for calories, a large peak at 110 and some small peaks; for fibre, the distribution focuses at 0-5, which might lead to high-variance, unstable prediction and overfitting.



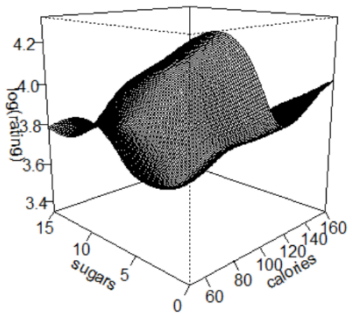
Question1: Based on the distribution of recommended calorie intake, are there any valuable relationship between sugar and rating.

According to National Research Council (1989), in your diet, no more than 10% of your calories should be consumed from simple carbohydrates (sugars), and no more than 30% should come from fat. Compared with other staple foods, cereals have lower fat content which determined by the nature of cereal (British Nutrition Foundation Nutrition Bulletin, p.21). But Sugar is artificially added later for taste. Hence, we set up GAM model to analysis correlation between sugar, rating and calories. From the correlation diagram above, it indicates that sugar and calories have a significant negative correlation to rating. Hence, we fit the GAM model to further analyse the relationship between these three variables.

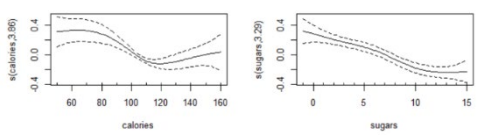
```r
cereals.gam <- mgcv::gam(log(rating) ~ s(calories) + s(sugars), data = cereals)
summary(cereals.gam)
```

```
Family: gaussian
Link function: identity

Formula:
log(rating) ~ s(calories) + s(sugars)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.70170    0.01895   195.4   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
             edf Ref.df      F  p-value
s(calories) 3.862  4.756  6.763 4.62e-05 ***
s(sugars)   3.292  4.064 14.765 2.51e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.738   Deviance explained = 76.3%
GCV = 0.030913  Scale est. = 0.027639  n = 77
```

The p-value of calories (4.62e-05), sugar (2.513-09) and R-squared value (0.738) are consistent with the conclusion as the linear model regarding the individual effects that both calories and sugars have a significant effect on rating.



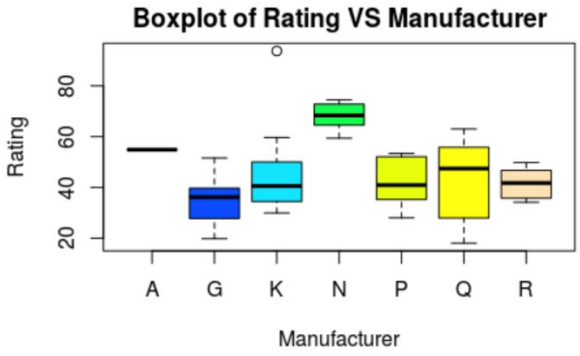

```r
par(mfrow = c(2, 2))
plot(cereals.gam)
```



From the generalized additive model and the 3D plot, we see the negative correlation that the rating decreases as sugar increases.

The rating decreases as calories increase at the beginning, however, when calories is 110, rating reaches the lowest point then it slightly increases as calories increases, it implies two extreme level of calories are more likely to have a higher rating.

Question2: Are there any breakfast cereals in market are disparate or virtually identical? Do manufactures have different strategies when adding depth of product line?



Comparing the median and range of each manufacturer, we notice that the mean rating of manufacturer Nabisco is the highest which scores around 67, then the second highest mean rating is from American Home Food Product. However, there is only one cereal (Maypo) from this manufacturer, and we can say the result might be not accurate and the sample needs more data from American Home Food product. The sample should be evenly collected from various manufacturers to avoid such bias.

The range of ratings from Quaker Oats is the most widespread, and its median is approximately 45, followed by Kelloggs, Post and Ralston Purina, their ratings are close, ranging from 40 to 50. For those customers with high-standard requirements for nutritional contents, Nabisco could be their fist choice. Or they can also choose the substitute products from Post, Kelloggs and Ralston for products from Quaker Oats.

Principle Component Analysis (question2 const.)

To illustrate the true dimensionality of this dataset, we draw the PCA graph using R packages include factoextra and ggpubr.

Accounting for over 90% of accumulated variability, there are six principal components to be considered. PC1 and PC2 contribute approximately 52% to the total variability, therefore simply reducing variable numbers is improper, which might cause missing useful information. Higher weights for those both in positive and negative direction, indicate whether the variables are correlated and how much correlation between variables.

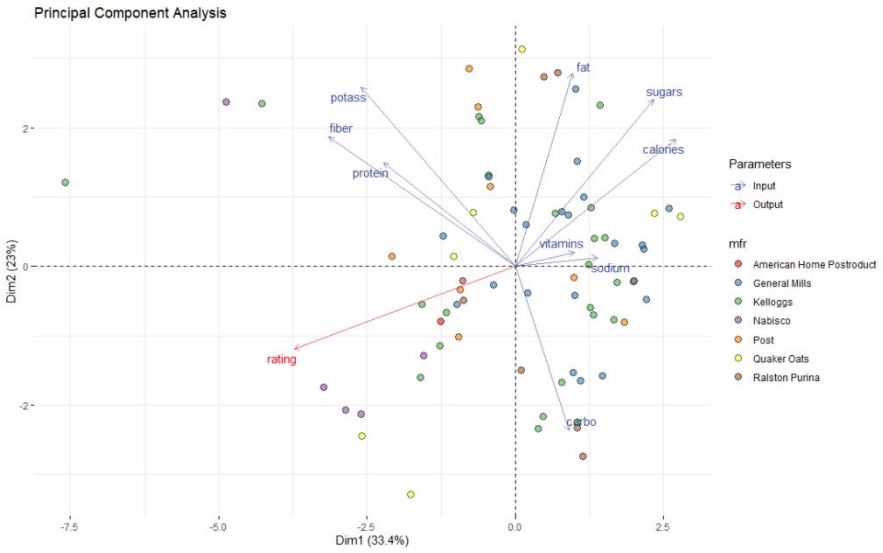
We look at the weights of each principal component for every variable, for instance, for the first principal component, fiber has negative weights, and calories and sugars have high positive weights. To be more precise, this means that for the cereal is low on fiber and high on calories and sugars, they associated with a high rating, its weight is around 0.55. When the negative weights increase, the positive weights would decrease correspondingly, they impact on each other. Seeing the second principal component, fibre, protein, fat and vitamins have high positive weights while the negative weights of carbohydrate is large. Continually, the third principal component is mainly associated with carbohydrate, protein, sodium and potassium; high scores on the fourth principal component is mainly associated with protein, fat and potassium; the fifth corresponds to factors including sodium, vitamins and fibre; the sixth has high positive weights on sugars and high negative weights on fat and sodium.

```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5     PC6     PC7
Standard deviation     1.7602 1.4328 1.2937 1.0043 0.9349 0.74486 0.65758
Proportion of Variance 0.3098 0.2053 0.1674 0.1009 0.0874 0.05548 0.04324
Cumulative Proportion  0.3098 0.5151 0.6825 0.7833 0.8708 0.92624 0.96948
                          PC8     PC9    PC10
Standard deviation     0.53067 0.14650 0.04644
Proportion of Variance 0.02816 0.00215 0.00022
Cumulative Proportion  0.99764 0.99978 1.00000
Standard deviations (1, .., p=10):
 [1] 1.76016816 1.43280849 1.29374567 1.00428702 0.93489751 0.74486336
 [7] 0.65757715 0.53067425 0.14650486 0.04643561
```

```
Rotation (n x k) = (10 x 10):
                PC1         PC2         PC3         PC4         PC5
Calories  0.457018731  0.07742639 -0.24681623  0.32049221 -0.10782665
Protein  -0.225456314  0.30800713 -0.42457827  0.40589453  0.14804590
Fat       0.262542858  0.42246286 -0.05758341  0.44663806 -0.16524346
Sodium    0.223843439 -0.18111381 -0.41066889 -0.05865237  0.62336151
Fibre    -0.310244865  0.42812506 -0.17065796 -0.12182832  0.38595823
Carbo     0.007058576 -0.50780287 -0.41643998  0.20912210 -0.29102617
Sugar     0.429026911  0.25345501  0.25260665 -0.17231182  0.20436158
Potass    0.174569599 -0.01457190 -0.49662713 -0.55288795 -0.07787114
Vitamins  0.005921914  0.42662995 -0.25393365 -0.36540484 -0.51440215
Rating   -0.557129508  0.01549073 -0.09681390  0.04623184 -0.06688595
                PC6         PC7         PC8         PC9         PC10
Calories  0.34241353 -0.32281102  0.16835314 -0.55712949  0.222910093
Protein   0.32712299  0.18848275 -0.55153304  0.10781785 -0.170346153
Fat      -0.41136215  0.36666223  0.35611506  0.27872322  0.119687046
Sodium   -0.49202854 -0.14855422 -0.16890631  0.01934191  0.251712933
Fibre     0.07470086 -0.31111895  0.55384473 -0.00827357 -0.344004912
Carbo     0.07673086 -0.30602813  0.21315029  0.50207716 -0.207799236
Sugar     0.40875834 -0.22645740 -0.09128263  0.58189782  0.226113676
Potass    0.26529410  0.54547364  0.19763386 -0.01029063  0.061952556
Vitamins -0.32127590 -0.40010708 -0.30303974  0.01191031  0.007773868
Rating    0.11082522 -0.06632806  0.14394262  0.09443806  0.791873122
```

The PCA plot illustrate the cereal from different manufacturer. Arrows in the first quadrant (fiber, potassium, protein) are positive indicators and arrows in fourth quadrant are negative indicators. According to this graph, most of the products are similar. Note that products from Nablsco and Quaker Oats is different from the others. The purple and yellow dots (represents cereals from Nablsco and Quaker Oats) in 2nd and 3th quadrant are healthier since they locate close to the arrow of potass, protein and potass, these are positively correlate to rating. In conclusion, cereals that are segmented into health food products have a more differentiated nutritional composition, which can be better adapted to the nutritional needs of different populations. And the product in 3rd and 4th quadrant are usually virtually identical which may deign for better tasty (more sugar).

As we mentioned in question2, most product from Nablsco are nutritious as it has the highest mean rating. The Quaker Oats have depth product line that this brand can cover nearly all your different needs for cereal. And we



could find out that Kelloggs and Post have similar product line depth. General Mills focuses more on the taste than the healthiness of the cereal.

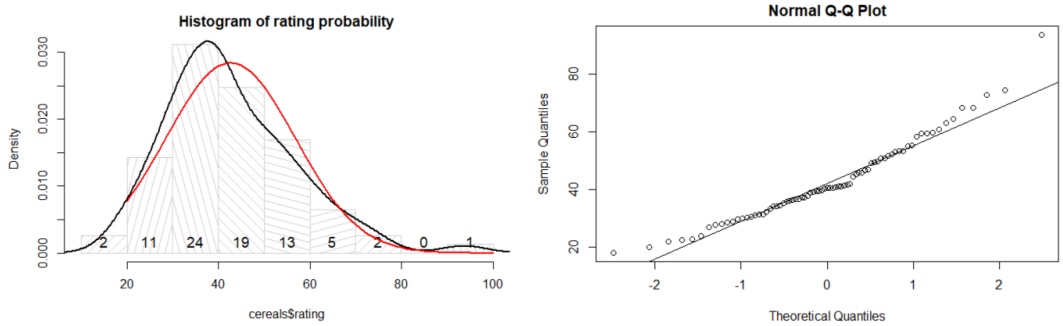
Question3: Are most of the cereals in the market are healthy that can be a good choice for those trying to lose weight?

```
        One Sample t-test

data:  cereals$rating
t = -4.4939, df = 67, p-value = 1.419e-05
alternative hypothesis: true mean is less than 50
95 percent confidence interval:
     -Inf 45.19928
sample estimates:
mean of x
 42.36582
```

Cereal is always considered as a healthy meal. To check if cereal is healthy at nutritional content point of view, we construct a hypothesis testing for rating.

Null Hypothesis: mean of rating >=50 vs Alternative Hypothesis: mean of rating <50

The P-value here is 1.49e-05 < 0.05, so we reject Null Hypothesis. We are 95% percent confident that rating is low than 45 while the mean is 42.6657. The result is surprising since cereal turns out to be not as health as we expect. Later, we need to do more analysis on it by simulations.



The left graph shows the histogram of rating probability. The red curve is the normal density curve that we expected, and the black curve is the actual density curve. And we could notice that the black curve does not fit the red curve well. From the normal QQ plot, we could find that the line fits the data points well, but the angle of line is not 45 degree. Both two graph means the rating is not a normal distribution, it is right skewing. The ratings for most of the cereals are under 50. It is not a high score at nutritional point of view.

The Bootstrap (follow Q3)

We want to do a simulation using the data from rating.

Bootstrapping method will be introduced in Chapter 11 lecture notes, which is the content for next week. But in order to use simulation strategies, we learned from lecture notes and online resources by ourselves. So we might make some unclear points (the code, or the interpretations). Please bear with us, thank you!

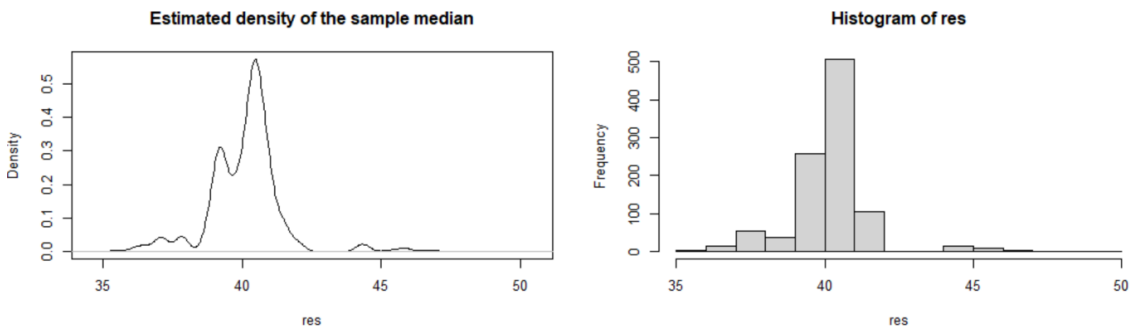
We are using Bootstrap as our simulation strategies to investigate the accuracies of our estimators and estimate the sampling distribution.

It can also provide us a result that we should reject or fail to reject the hypothesis testing that:

Null hypothesis: mean rating > 50 vs Alternative hypothesis: mean rating <= 50

Step 1:

We simulate 1000 resamples from the empirical distribution function (rating). Then we draw a histogram for those resamples. The bootstrap distribution is the distribution of means from each resample. The bootstrap



distribution should appear to be normal. If the bootstrap distribution is non-normal, we cannot trust the results. We also draw a density estimate for those resamples.

Step 2:

We examine the confidence interval by using the boot function in the "boot" library. The mean of the bootstrap sample is an estimate of the population mean, which is the mean of rating. Because the mean is based on sample data (rating of 74 obs in our dataset) and not the entire population, it is unlikely that the sample mean equals the population mean.

In order to estimate the population mean better, we use the confidence interval. Confidence intervals are based on the sampling distribution of a statistic. If a statistic has no bias as an estimator of a parameter, its sampling distribution is centered at the true value of the parameter. A bootstrapping distribution approximates the sampling distribution of the statistic. Therefore, the middle 95% of values from the bootstrapping distribution provide a 95% confidence interval for the parameter. The confidence interval helps us assess the practical significance of our estimate for the population parameter.

Interpretations:

From the histogram and the density estimate of our 1000 resamples, we notice that the population mean (mean rating) is around 40. There is an issue with the bootstrap distribution. We can see that the bootstrap distribution is non-normal, which means we cannot trust the results entirely.

From the R output, based on 1000 bootstrap replicates, we have:

95% Confidence Interval (38.80, 43.66).

In these results, the estimate for the population mean is approximately ---. You can be 95% confident that the population mean is between approximately 38.80 and 43.66.

```
> boot.ci (rating.bt, conf = 0.95, type = "basic")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = rating.bt, conf = 0.95, type = "basic")

Intervals :
Level      Basic
95%   (38.80, 43.66 )
Calculations and Intervals on Original Scale
```

Hence, we can reject the null hypothesis that the mean rating is larger than 50.

The bootstrap method uses a very different approach to estimate sampling distributions. This method takes the sample data that a study obtains, and then resamples it over and over to create many simulated samples. Each of these simulated samples has its own properties, such as the mean. When you graph the distribution of these means on a histogram, you can observe the sampling distribution of the mean. You don't need to worry about test statistics, formulas, and assumptions.

The bootstrap procedure uses these sampling distributions as the foundation for confidence intervals and hypothesis testing. Let's take a look at how this resampling process works.

Assessment

This model still has some limitations, now we will states them and provide ways to improve our analysis.

1. This dataset focuses on the nutritional contents, but for a complete and precise dataset, it is essential to have some non-nutritional factors for us to analyse. The result of our analysis could be more useful if we have more predictors in the dataset for the companies point of view. For example, by the business aspect, we can explore the customer's satisfaction, the sales quantity, price of the product or even the packaging. If the data is not limiting and we have more predictors in hand, we can analyse and know which attributes of cereals that customers value more in their purchases. It is a good advice for manufacturers to know that which aspects that they should pay more attention on to make their product more popular and boost the sales.

2. Linear regression model is sensitive to outliers and data should be independent. Linear regression could only tell us the linear relationship, but this cannot tell the whole story for this dataset most of the time. It may over simplifies many real-world problems.

3. GLMs, GAMs and so on rely on assumptions about the data generating process. If those are violated, the interpretation of the weights is no longer valid (Christoph, 2020). Hence, we should make sure the hypothesis holds in advance to fitting a model.

4. The data is record in 1993. The nutritional composition of the cereal at each manufactural will change dramatically over the course of 27 years. This data is not time sensitive. The conclusions we draw do not maximize the usefulness of our clients' purchase.

5. We only have data for one year and things may vary over years, so we are unable to observe changes in the breakfast cereal market in a time dimension. For instance, there was movement towards high fiber cereals from 1978 to 1987 (Pauline 1989, p63).

Conclusion

We fit the multiple linear regression model and it suggests that there is a linear relationship between the rating and these 9 nutritional variables include calories, sugars, protein, fat, sodium, fiber, carbo, potass, and vitamins.

Our analysis indicates that a cereal with high-fiber, high-protein, low-fat, and low-sugar would be more nutritious. This can be a reasonable criterion for us to choose a right healthy cereal. Hence, next time when you purchase in grocery store, it is a good idea to choose those with high fiber, high protein, low fat, and low sugar. Moreover, as we find the fact that Nabisco has significantly higher ratings than others since it produces cereals that high in fiber with no fat and sugar, choosing Nabisco can also a healthy choice for cereals.

To link our study to the real world, think of the marketing strategies and advertisements that businesses use to boost their sales. Nowadays, companies and manufacturer put large attention on emphasising the advantages of ingredients in their products. For example, they might add the words "Low Fat" "Sugar Free" or "With Extra Fiber" to give their products a healthy sounding name.

To summarize, we fit linear model, generalized additive model, and Bootstrap simulation and we draw graphs as data visualization to interpret them. In our report, we arise three interesting questions and answer them by providing statistical analysis to them. The result provides a criterion for customers to make purchase wisely. The cereal products on the cereal breakfast market have different characteristics, so customers can choose the right product that satisfies their own needs. Most product from Nablsco contains nutritious ingredient, while Quaker Oats have depth product line since this manufacturer produces various types of cereals that provides customers different needs for cereal. From question three and Bootstrap, we find that cereal is not that nutritious as we expected.

Reference

Christoph, M 2020, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, e-book, accessed 1 August 2020, <https://christophm.github.io/interpretable-ml-book/extend-lm.html>

Green, P., Silverman, B. 1984, Stochastic relaxation, gibbs distribution, and the Bayesian restoration of images, IEEE Trans, Pattern Anal, Mach, Intell,6:721-741.

National Research Council, 1989a. "Diet and Health: Implications for Reducing Chronic Disease Risk". National Academy Press, Washington, D.C.

Lafaye de Micheaux, P., Drouilhet, R., and Liquet, B. 2013, The R Software: Fundamentals of Programming and Statistics Analysis, Statistics and Computing. Springer New York.

Pauline M. Ippotito, Alan D. Mathios 1989, 'Healthy claims in advertising and labeling: A study of the Cereal Market', paper presented at FEDERAL TRADE COMMISSION, August.

Brigid, M 2004, 'Nutritional aspects of cereals', British Nutrition Foundation Nutrition Bulletin, No.29, pp.111-142, accessed 02 August 2020 from Sage Journals Online.

DataCamp. (2018, October 10). Bootstrap in R. Retrieved from https://www.datacamp.com/community/tutorials/bootstrap-r

Appendix

```
# import the cereal dataset
data <- read.csv(file.choose())
cereals <- data.frame(data)

# ---------- Data cleaning and wrangling----------
# delete three obseravtions as they have missing values
cereals <- cereals[-c(5, 21, 58), ]

# change abbrev manufacturer name to full name
cereals$mfr_full <- cereals$mfr
cereals$mfr_full <- gsub(pattern = "A", replacement = "American Home Product", x = cereals$mfr_full)
cereals$mfr_full <- gsub(pattern = "G", replacement = "General Mills", x = cereals$mfr_full)
cereals$mfr_full <- gsub(pattern = "K", replacement = "Kelloggs", x = cereals$mfr_full)
cereals$mfr_full <- gsub(pattern = "N", replacement = "Nabisco", x = cereals$mfr_full)
cereals$mfr_full <- gsub(pattern = "P", replacement = "Post", x = cereals$mfr_full)
cereals$mfr_full <- gsub(pattern = "Q", replacement = "Quaker Oats", x = cereals$mfr_full)
cereals$mfr_full <- gsub(pattern = "R", replacement = "Ralston Purina", x = cereals$mfr_full)

# change the type H and C to "Hot" and "Cold"
cereals$type <- gsub(pattern = "H", replacement = "Hot", x = cereals$type)
cereals$type <- gsub(pattern = "C", replacement = "Cold", x = cereals$type)

# change "type" and "shelf" to factor
cereals$type <- factor(cereals$type)
cereals$shelf <- factor(cereals$shelf)
sapply(cereals, FUN = class)
summary(cereals)
#-------------------------------------------------------------------------
# ----------Exploratory Data Analysis----------

# Scatterplots for rating and 9 exploratory variables
# Which nutrients are essential for a nutritious breakfast per rating?
# In order to put these 9 plots all together
par(mfcol=c(3,3))
# rating vs calories
```

```
plot(rating~calories,
    data = cereals,
    xlab="Calories",
    ylab="Rating",
    main="Rating vs Calories",
    col="green")
abline(lm(cereals$rating~cereals$calories), col="red")

# rating vs protein
plot(rating~protein,
    data = cereals,
    xlab="Protein",
    ylab="Rating",
    main="Rating vs Protein",
    col="green")
abline(lm(cereals$rating~cereals$protein), col="red")

# rating vs fat
plot(rating~fat,
    data = cereals,
    xlab="Fat",
    ylab="Rating",
    main="Rating vs Fat",
    col="green")
abline(lm(cereals$rating~cereals$fat), col="red")

# rating vs sodium
plot(rating~sodium,
    data = cereals,
    xlab="Sodium",
    ylab="Rating",
    main="Rating vs Sodium",
    col="green")
abline(lm(cereals$rating~cereals$sodium), col="red")

# rating vs fiber
```

```
plot(rating~fiber,
    data = cereals,
    xlab="Fiber",
    ylab="Rating",
    main="Rating vs Fiber",
    col="green")
abline(lm(cereals$rating~cereals$fiber), col="red")

# rating vs carbo
plot(rating~carbo,
    data = cereals,
    xlab="Carbohydrates",
    ylab="Rating",
    main="Rating vs Carbohydrates",
    col="green")
abline(lm(cereals$rating~cereals$carbo), col="red")

#rating vs sugars

plot(rating~sugars,
    data = cereals,
    xlab="Sugar",
    ylab="Rating",
    main="Rating vs Sugar",
    col="green")
abline(lm(cereals$rating~cereals$sugars), col="red")

# rating vs potass
plot(rating~potass,
    data = cereals,
    xlab="Potassium",
    ylab="Rating",
    main="Rating vs Potassium",
    col="green")
abline(lm(cereals$rating~cereals$potass), col="red")
```

```
# rating vs vitamins
# note that we only have 0, 25 and 100 as the value of vitamins
# hence it is better to do a boxplot for vitamins
boxplot(rating~vitamins,
        data = cereals,
        xlab="Vitamins",
        ylab="Ratings",
        main="Rating vs Vitamins",
        col=c("red","green","blue"))

# ------------------------------------------------------------------------------
# ----------------------------------------------Model Fitting-----------------------------------
# fit a full model with all variables first
fit_full <- lm(rating~mfr+type+calories+protein+fat+sodium+fiber+sugars+
                 potass+vitamins+shelf+weight+cups, data = cereals)
summary(fit_full)

# after dropping some un-significant predictors, we now use
# AIC forward selection
library(MASS)
fit_testing <- lm(rating~calories+protein+fat+sodium+fiber+carbo+
                    sugars+potass+vitamins, data = cereals)
fit_initial <- lm(rating~1, data = cereals)
AIC.model<-stepAIC(fit_initial, formula(fit_testing), direction=c("forward"))
AIC.model_summary<-summary(AIC.model)
AIC.model_summary

# by the output, we get the final model
fit_final <- lm(rating~sugars+fiber+sodium+fat+protein+vitamins, data=cereals)

# Diagnostics for the final model
par(mfrow=c(2,2))
plot(fit_final)

# ------------------------------------------------------------------------------
# ---------------------------General Additive Model-----------------------------------
```

```
# fitting a GAM and plot the graphs for it
cereals.gam <- mgcv::gam(log(rating) ~ s(sugars) + s(calories), data = cereals)
par(mfrow = c(2, 2))
plot(cereals.gam)

# 3 dimensional plot for this model
grid <- list(calories = seq(from = 50, to = 160, length = 100),
sugars = seq(from = 0, to = 15, length = 100))
cereals.pr <- mgcv::predict.gam(cereals.gam, newdata = expand.grid(grid))
cereals.pr <- matrix(cereals.pr,ncol = 100, nrow = 100)
persp(grid$calories, grid$sugars, cereals.pr,
xlab = "calories", ylab = "sugars", zlab = "log(rating)",
theta = -45, phi = 15, d = 2.0, tick = "detailed")

# ------------------------------------------------------------------------------
# ---------------------------------Kernel Density-----------------------------------------
# kernel density for potass, fiber, sodium, vitamins and calories
par(mfrow=c(2,3))
kd.calories<-density(cereals$Calories)
plot(kd.calories, main = "Kernel Density of Calories")
polygon(kd.calories, col="blue", border = "black")

kd.potass<-density(cereals$Potass)
plot(kd.potass,main="Kernel Density of Potass")
polygon(kd.potass,col="blue",border = "black")

kd.fibre<-density(cereals$Fibre)
plot(kd.fibre,main="Kernel Density of Fibre")
polygon(kd.fibre,col="blue",border = "black")

kd.sodium<-density(cereals$Sodium)
plot(kd.sodium,main="Kernel Density of Sodium")
polygon(kd.sodium,col="blue",border = "black")

kd.vitamins<-density(cereals$Vitamins)
plot(kd.vitamins,main="Kernel Density of Vitamins")
```

```
polygon(kd.vitamins,col="blue",border = "black")


kd.fat<-density(cereals$Fat)
# Is there any significance of manufacturer with regards to cereal rating?
boxplot(rating~mfr,
        data = cp,
        xlab = "Manufacturer",
        ylab = "Ratings",
        main = "Rating vs Manufacturer",
        col = topo.colors(7))


# ------------------------------------------------------------------------------------------------
# ----------------------------------------------PCA analysis--------------------------------------
# use the functions from these two R packages
library(factoextra)
library(ggpubr)

# select out the variables that we want to analyse
PCA_cereals <- cereals %>%
  select(name, mfr_full, calories, protein, fat, sodium, fiber, carbo,
         sugars, potass, vitamins, rating)
PCA_cereals <- PCA_cereals[complete.cases(PCA_cereals),]
PCA_data <- prcomp(PCA_cereals[, 3:12], scale. = TRUE)
# get exact information of each PC group
summary(PCA_data)
report(PCA_data)

# draw the PCA graph for data visualization
fviz_pca_biplot(PCA_data,
                geom.ind = "point",
                pointshape = 21,
                pointsize = 3,
                fill.ind = PCA_cereals$mfr,
                alpha = 0.6,
                mean.point = FALSE,
                col.var = factor(c("Input", "Input", "Input", "Input", "Input", "Input", "Input", "Input", "Input", "Output")),
```

```
            repel = TRUE,
            legend.title = list(fill = "mfr", color = "Parameters"),
            title = "Principal Component Analysis") +
  fill_palette("Set1") +
  color_palette(palette = "aaas")

# ------------------------------------------------------------------------------------------------------
# ----------------------------------Hypothesis Testing for rating-----------------------------------
t.test(cereals$rating,alternative="less",mu=50)

hrp<-hist(cereals$rating,freq=FALSE,density=10,angle=15+30*1:6,main="Histogram of rating probability")
text(hrp$mids,0,hrp$counts,adj=c(.5,-.5),cex=1.2)
lines(density(cereals$rating),lwd=2)
p<-seq(from=20,to=100,by=1)
lines(p,dnorm(p,mean(cereals$rating),sd(cereals$rating)),col="red",lwd=2)

qqnorm(cereals$rating)
qqline(cereals$rating)

# ------------------------------------------------------------------------------------------------------
#-----------------------------------------------Bootstrap-----------------------------------------------

# We simulate three samples from the empirical distribution function F_hat
# Notice how the resamples may contain repeats of values in the original sample

# assign a sample to store the sample data
# simulations for rating values
asample <- cereals$rating

# construct a loop to get 1000 resamples
res <- rep(0, times = 1000)
for (i in 1:1000) {
  res[i] <- median(sample(asample, replace = TRUE))
}

# draw a histogram of our 1000 resamples
```

```
hist(res)

# density estimate for the vector res
plot(density(res), xlab = "res", ylab = "Density", type = "l", main = "Estimated density of the sample median")

library("boot")
# Rescale the data
rating <- asample
rating.bt <- boot(data = rating, statistic = function(d, i)
  median(d[i]), R = 1000)

# Here rating contains the data, median is the function which computes the
# statistic of interest (the sample median), and R is the
# number of boostrap samples. The function median() can be replaced by
# some other (possibly user defined) function.

# 95% confidence interval for our resamples
boot.ci (rating.bt, conf = 0.95, type = "basic")
```