Deriving a Simplified Multilinear Model to Determining Cereal Nutrition Ratings

STAT 481 Project 1

By: Andrew Gascon

March 29, 2024

Abstract

The purpose of this study is to create a simplified multiple linear regression model which performs comparatively well at determining the nutritional rating of a cereal brand compared to its more complex fully constructed counterpart. This simplified model would ideally be easier to interpret, alleviate the issue of overfitting that more complex models suffer from, and still perform comparatively well to their more complex counterpart. First, we introduce and summarize the dataset, then we construct a full multilinear model by evaluating its model assumptions, and then using backward model selection to create a simplified multilinear model. Both models are evaluated against their assumptions and their performances are compared to each other using their respective R^2 scores, which measures what percentage of the variability in the response variable can be explained by the model's predictive variables. Of the original 7 predictive variables used by the initial fully constructed multilinear model, the final simplified model uses just 4, but still performs similarly well to its more complex counterpart. We find that when other variables are held constant, milligrams of sodium and grams of sugar have an adverse effect on a cereal brand's health rating, whereas grams of dietary fiber and number of cups of one serving have a positive impact. Our final, simplified model has an adjusted R^2 of 0.8968, which shows that it still does an excellent job at determining the nutritional rating of cereal brands, even compared to the initial fully constructed model.

Data Introduction and Description

Data Introduction:

The dataset under analysis contains 14 variables consisting of nutritional facts and other observable characteristics for 77 cereal brands. Of these 14 variables, 8 will be used for the regression models: protein, sodium, fiber, sugars, vitamins, weight, cups, and rating. The first 7 of these listed variables will serve as predictor variables and the last of these listed variables, rating, will serve as the response variable. The rating variable, calculated by Consumer Reports, is used to quantify how nutritious a particular brand of cereal is. A full list of variables and a brief description is provided below:

Variable Name	Description
Name	Name of cereal
type	cold or hot
calories	calories per serving
protein (x_1)	grams of protein
fat	grams of fat
sodium (x_2)	milligrams of sodium
fiber (x_3)	grams of dietary fiber
carbo	grams of complex carbohydrates
sugars (x_4)	grams of sugars
potass	milligrams of potassium
vitamins (x_5)	vitamins and minerals - 0, 25, or 100, indicating the typical percentage of FDA recommended
shelf	display shelf (1, 2, or 3, counting from the floor)
weight (x_6)	weight in ounces of one serving
cups (x_7)	number of cups in one serving
rating (y)	a rating of the cereals (response)

The purpose of this study is to derive a simplified multiple linear regression model that still performs comparatively well at determining the nutrition rating of a brand of cereal compared to its more complex fully constructed counterpart. The multiple linear regression model is defined in the form:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} + \epsilon_i$$

$$i = 1, 2, \dots, n, where\ \epsilon_i = N(0, \sigma^2)$$

Where ϵ_i are independent and identically distributed (iid)

In this model, the variables are defined as follows:

- Y_i = rating (as calculated by Consumer Reports)
- x_1 = proteins (grams)
- x_2 = sodium (milligrams)
- x_3 = dietary fiber (grams)
- x_4 = sugars (grams)
- x_5 = vitamins and minerals
- x_6 = weight (ounces per serving)
- x_7 = cups per serving

Descriptive Statistics:

Before creating the linear models, we must first provide a summary of the variables of interest to have a better understanding of the dataset and to identify any problems that may affect model performance, such as outliers. Summaries of each variable are given by their five-number summary (minimum, lower/first quartile, median, upper/third quartile, and maximum), mean, standard deviation, and number of outliers. The distribution of each variable is visualized using a histogram and a boxplot, the latter of which visualizes their outliers. The summary is as follows:

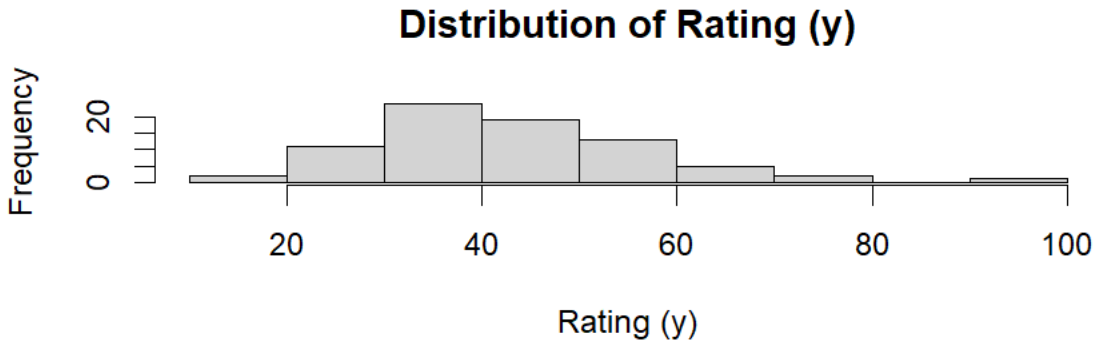
Y (rating):

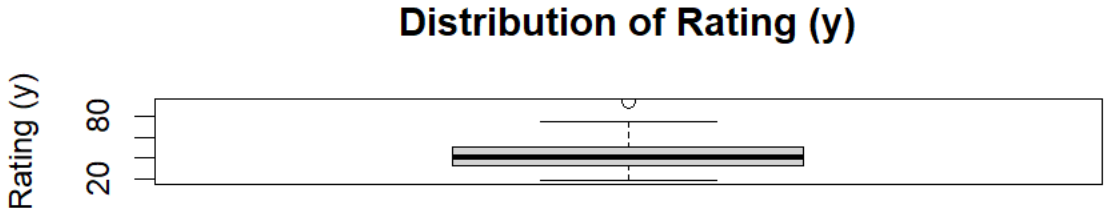
This variable has a large range (75.66206) and the mean (42.6657) is close to the median (40.40021).

Minimum	Lower/First Quartile	Median	Upper/Third Quartile	Maximum	Mean	Standard Deviation	Number of Outliers
18.04285	33.17049	40.40021	50.82839	93.70491	42.6657	14.04729	1

Histogram and Boxplot:

The distribution of Y (rating) is skewed right and has one outlier.





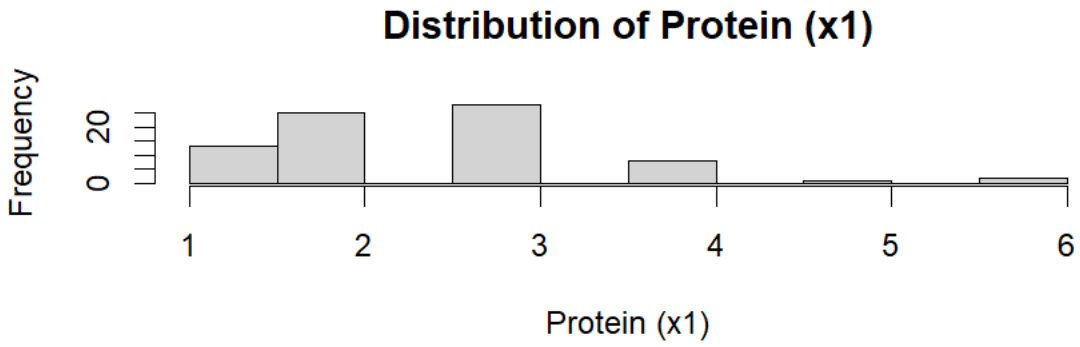
x1 (proteins):

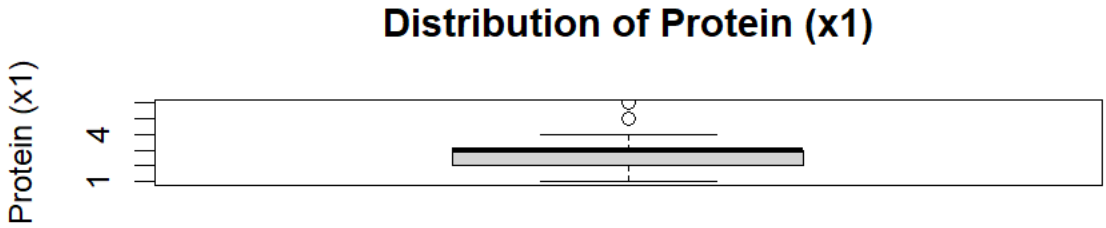
This variable has a small range (5) and the mean (2.545455) and the median (3) are close.

Minimum	Lower/First Quartile	Median	Upper/Third Quartile	Maximum	Mean	Standard Deviation	Number of Outliers
1	2	3	3	6	2.545455	1.09479	3

Histogram and Boxplot:

The distribution of x1 (proteins) is skewed left and has 3 outliers.





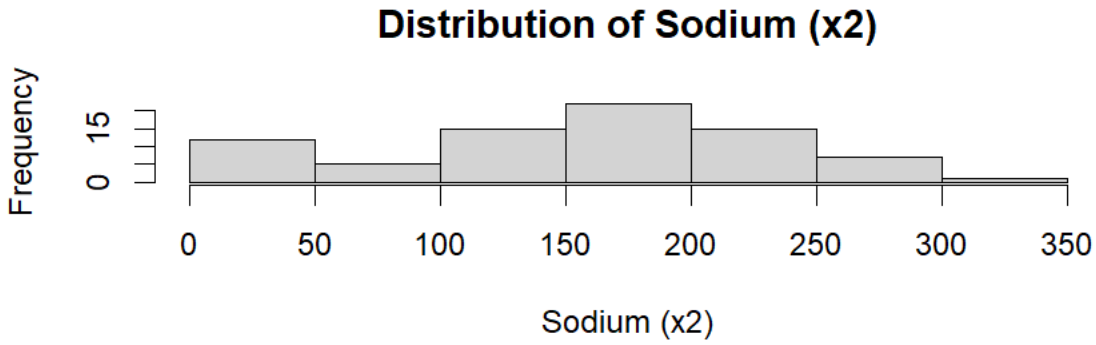
x2 (sodium):

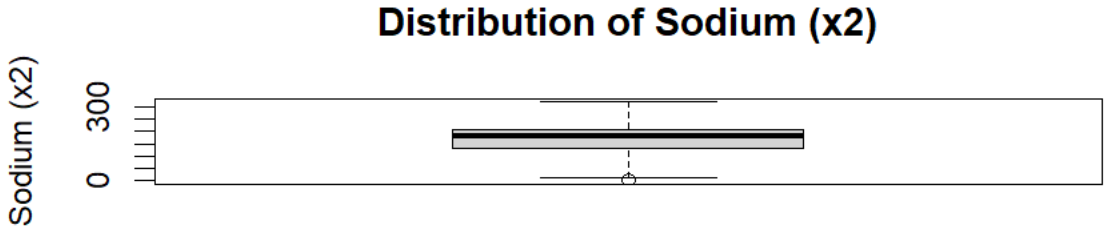
This variable has a large range (32). There is a large distance between the mean (159.6753) and the median (180).

Minimum	Lower/First Quartile	Median	Upper/Third Quartile	Maximum	Mean	Standard Deviation	Number of Outliers
0	130	180	210	320	159.6753	83.8323	9

Histogram and Boxplot:

The distribution of x2 (sodium) is slightly bimodal and has 9 outliers.





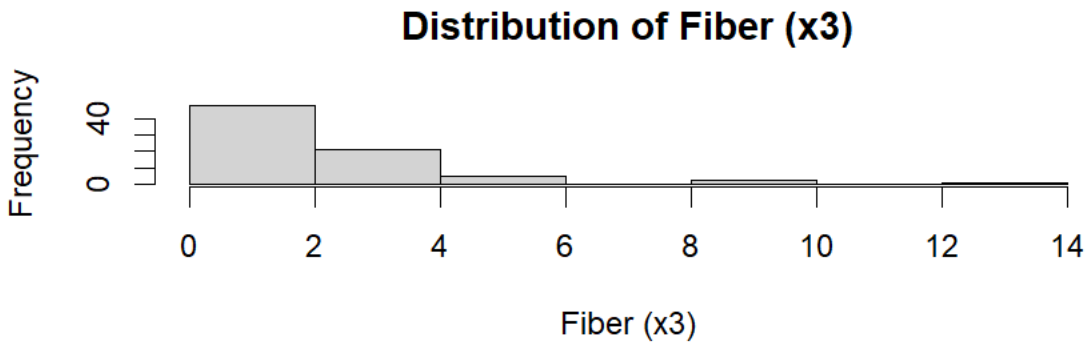
x3 (dietary fiber):

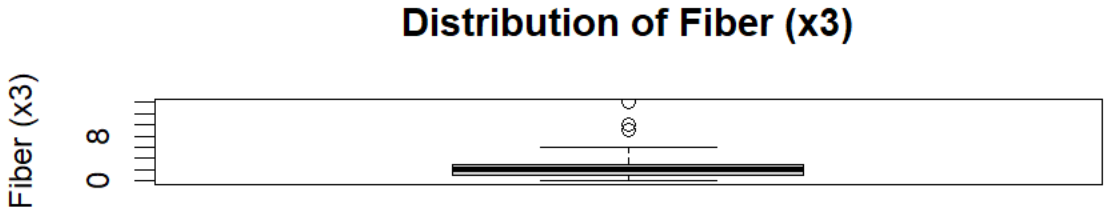
This variable has a large range (14) and the mean (2.151948) and the median (2) are close.

Minimum	Lower/First Quartile	Median	Upper/Third Quartile	Maximum	Mean	Standard Deviation	Number of Outliers
0	1	2	3	14	2.151948	2.383364	3

Histogram and Boxplot:

The distribution of x3 (dietary fiber) is skewed left and has 3 outliers.





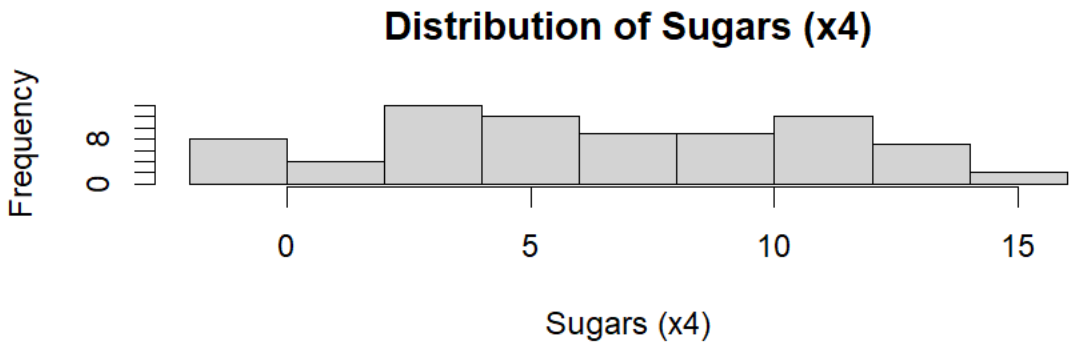
x4 (sugars):

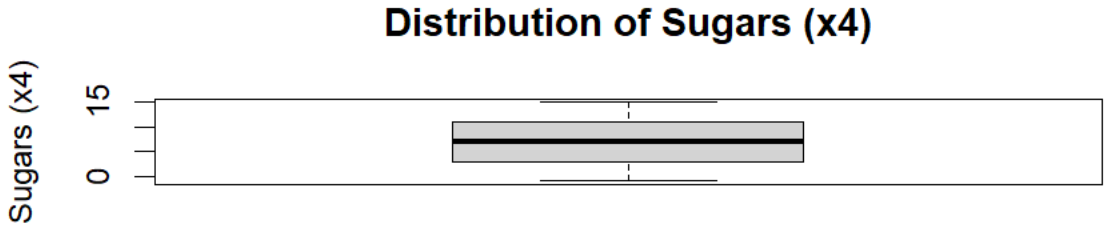
This variable has a large range (16) and the mean (6.922078) and the median (7) are close.

Minimum	Lower/First Quartile	Median	Upper/Third Quartile	Maximum	Mean	Standard Deviation	Number of Outliers
-1	3	7	11	15	6.922078	4.444885	0

Histogram and Boxplot:

The distribution of x4 (sugars) is bimodal and has no outliers.





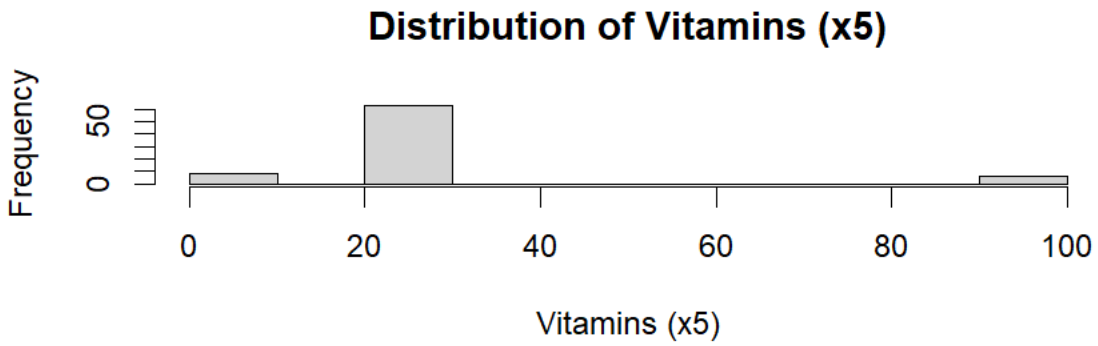
x5 (vitamins and minerals):

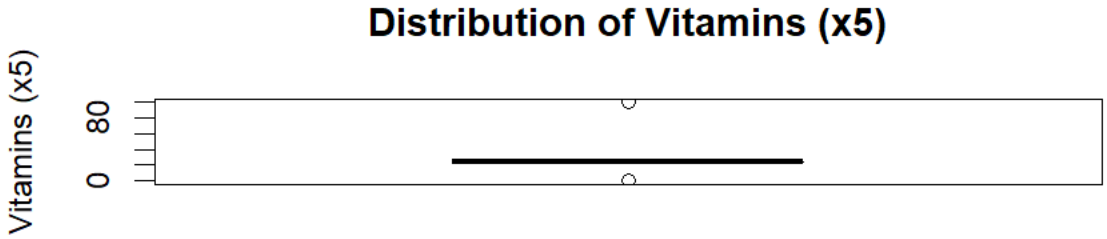
This variable has a large range (100) and the mean (28.24675) and the median (25) are close.

Minimum	Lower/First Quartile	Median	Upper/Third Quartile	Maximum	Mean	Standard Deviation	Number of Outliers
0	25	25	25	100	28.24675	22.34252	14

Histogram and Boxplot:

The distribution of x5 (vitamins and mineral is skewed right and has 14 outliers.





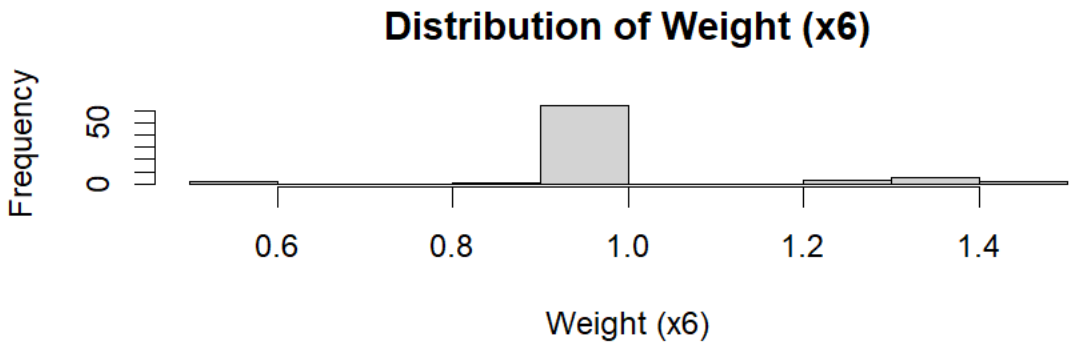
x6 (weight):

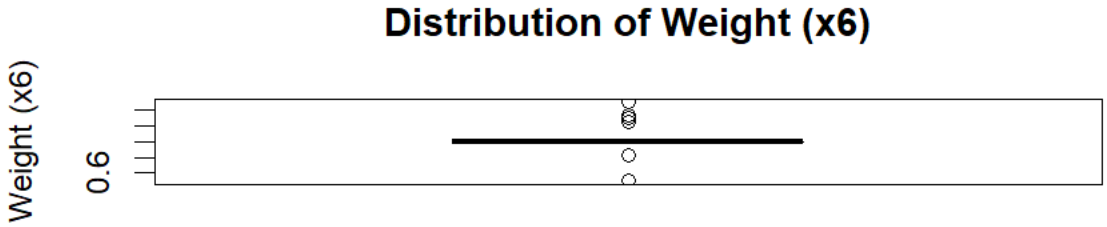
This variable has a small range (1) and the mean (1.02961) is close to the median (1).

Minimum	Lower/First Quartile	Median	Upper/Third Quartile	Maximum	Mean	Standard Deviation	Number of Outliers
0.5	1	1	1	1.5	1.02961	0.1504768	13

Histogram and Boxplot:

The distribution of x6 (weight) is very unimodal but has 13 outliers.





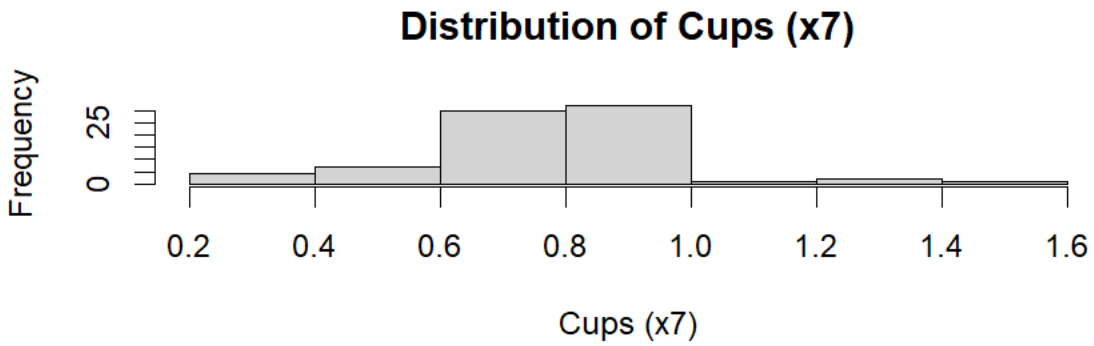
x7 (cups per serving):

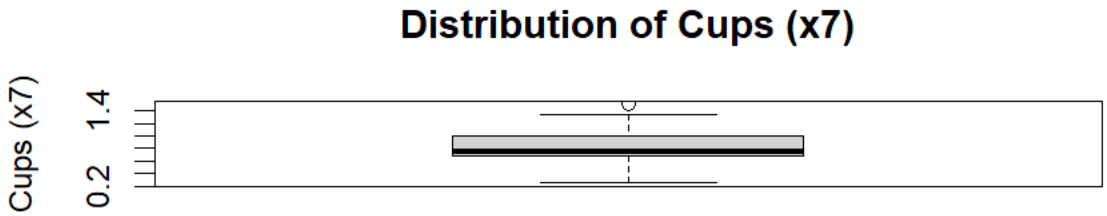
This variable has a small range (1.25) and the mean (0.821039) is close to the median (0.75)

Minimum	Lower/First Quartile	Median	Upper/Third Quartile	Maximum	Mean	Standard Deviation	Number of Outliers
0.25	0.67	0.75	1.00	1.50	0.821039	0.2327161	1

Histogram and Boxplot:

The distribution of x7 (cups per serving) is unimodal and has 1 outlier.





Multiple Linear Regression Analysis

Multicollinearity:

Before we proceed with the multiple linear regression analysis, we must run a check for multicollinearity within the predictor variables to ensure that there are no correlations between the predictor variables themselves. To check for this, we calculate the Variance Inflation Factor (VIF) for each predictor variable. If the VIF for a variable is greater than 10, then the multicollinearity of that variable may be influencing the estimates of other variables, thus necessitating their removal from the model. We only need to run this check once at the beginning of the analysis. Per the R code, all the predictor variables have a VIF less than 10, meaning there is no significant multicollinearity.

```
> # Multicolinearity check
> # Full model
> model <- lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7, data = analysis)
> library(car)
Loading required package: carData
> vif(model)
      x1       x2       x3       x4       x5       x6       x7
1.656329 1.261257 1.748384 1.701450 1.250893 1.938023 1.429315
```

Model Assumptions:

Now that we've assured that there is no significant multicollinearity present that may disrupt the rest of this analysis, we are going to construct the regression model of the form:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} + \epsilon_i$$

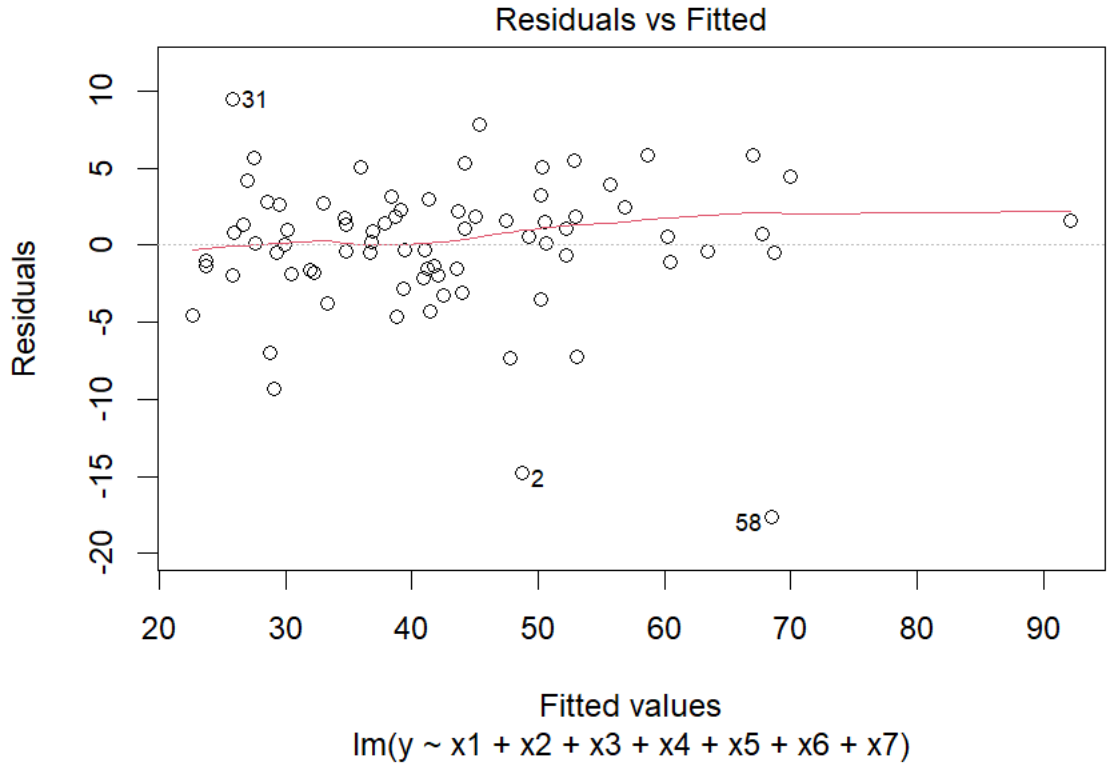
$$i = 1, 2, \dots, n, where\ \epsilon_i = N(0, \sigma^2)$$

Where ϵ_i are independent and identically distributed (iid)

Next, we must verify that the fully constructed regression model meets the following four criteria: linearity of the model, independence of errors, normality of errors, and equal variance of errors. If any of these conditions are not met, then the identified corresponding variables will be transformed accordingly. If all of these criteria are met, then we will proceed with the analysis.

Linearity:

To check for linearity, we must plot Residuals vs. Fitted Values and look for any obvious patterns in the resulting scatterplots. The residual vs. fitted values plot indicates that variances are decently spread out throughout the plot. This suggests that the linearity condition holds.

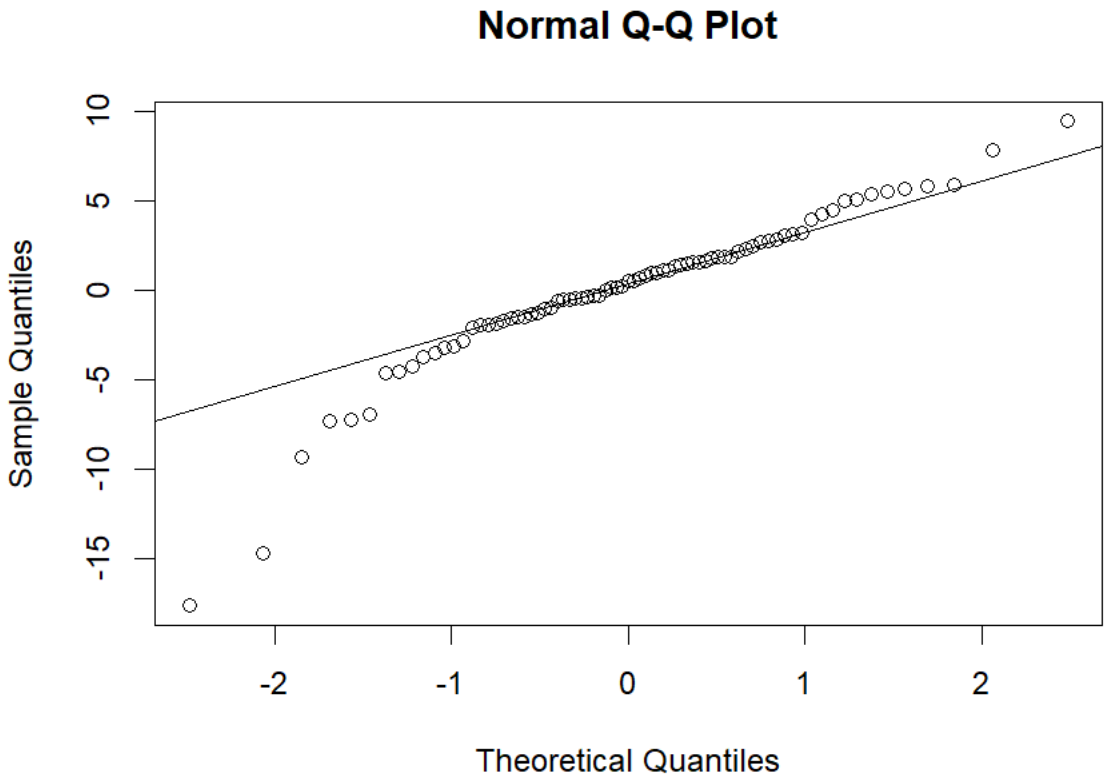


Independence of Errors:

Because our data is not time-series and was not collected in a sequence, we can assume that we have an independence of errors.

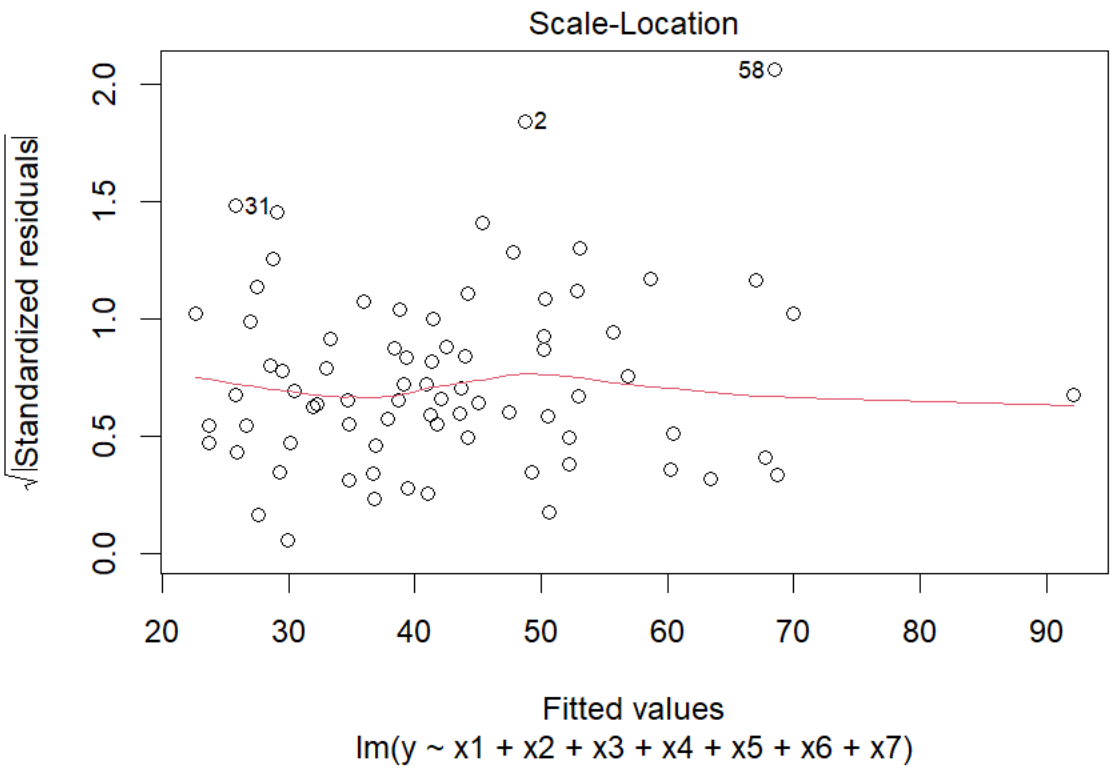
Normality of Errors:

Per the results of the QQ Plot, the distribution of residuals is mostly normal, especially near 0, with one tail (upper right corner) being slightly skewed. However, the other tail (bottom left corner) has a relatively small number of seriously skewed outliers. Due to the proportion of residuals that can be considered normally distributed, we can determine that the residuals are normally distributed, albeit with significant outliers.



Equal Variance of Errors:

To check for equal variances, we check the plot of standardized residuals vs fitted values to look for any obvious patterns as indicated by the red line. Because the red line shows no obvious patterns, we can assert that the equal variance of errors condition holds.



Because our model satisfies all four criteria, we can proceed to the next step of the analysis.

Variable Selection

Primary Test Hypothesis and Partial t-tests

Our model is currently in the form:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} + \epsilon_i$$

$$i = 1, 2, \dots, n, where\ \epsilon_i = N(0, \sigma^2)$$

Where ϵ_i are independent and identically distributed (iid)

Our next step is to identify which coefficient parameters are significant in explaining the variability in rating (Y). To do this, we use backward selection to remove any predictor variables that are not deemed significant, which is determined by whether their significance is 0.10 or lower. Backward selection is being used because our current, full model has many variables, meaning it has a higher complexity and thus a higher risk of overfitting. Backward selection aims to create a smaller, less complex model to alleviate this. After this process, we will re-run our model assumptions on this final model.

Below is a summary of our full model analysis of the coefficients which includes their estimate, standard error, t value, and p-value.

```
> summary(model)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7, data = analysis)

Residuals:
     Min       1Q   Median       3Q      Max
-17.6373  -1.5977   0.5309   2.2725   9.4545

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 56.494977   4.783487  11.810  < 2e-16 ***
x1           0.187655   0.614325   0.305   0.7609
x2          -0.047937   0.007001  -6.847 2.51e-09 ***
x3           3.011883   0.289923  10.389 9.47e-16 ***
x4          -2.006770   0.153357 -13.086  < 2e-16 ***
x5          -0.027392   0.026160  -1.047   0.2987
x6          -2.225473   4.834653  -0.460   0.6467
x7           4.655274   2.684682   1.734   0.0874 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.556 on 69 degrees of freedom
Multiple R-squared:  0.9045,    Adjusted R-squared:  0.8948
F-statistic: 93.37 on 7 and 69 DF,  p-value: < 2.2e-16
```

First, we will test whether any of our predictor variables can explain the variability in rating. This is done through our **Primary Test Hypothesis:**

$$H_0: \beta_1 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7$$

$$H_1: \text{at least one } \beta_j \neq 0 \text{ for j} = 1, 2, \ldots, 7$$

The p-value of our full model analysis (< 2.2e-16), is less than our significance value of 0.10, which means we can reject the null hypothesis, meaning that at least one of these predictive variables explains the variability in rating. The R^2 of our full model is 0.9045 with an adjusted R^2 of 0.8948. This means that 89.48% of the variability in rating can be explained by the full model. Next, we'll need to use a series of partial t-tests to determine which specific variables explain the variability in rating. These hypotheses are as follows:

$$H_0: \beta_j = 0 \text{ vs. } H_1: \beta_j \neq 0, \text{ for j} = 1, 2, \ldots, 7$$

According to the below ANOVA table of our full model, only x1 (protein), x2 (sodium), x3 (fiber), x4 (sugars), and x7 (cups) are significant in explaining the variability in rating. This is because their respective p-values are less than our significance threshold of 0.10.

```
> anova(model)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq  F value    Pr(>F)
x1         1 3321.5  3321.5 160.0340 < 2.2e-16 ***
x2         1 2121.6  2121.6 102.2218 2.961e-15 ***
x3         1 2215.6  2215.6 106.7480 1.194e-15 ***
x4         1 5815.3  5815.3 280.1888 < 2.2e-16 ***
x5         1   20.0    20.0   0.9627   0.32995
x6         1    8.3     8.3   0.3986   0.52990
x7         1   62.4    62.4   3.0068   0.08738 .
Residuals 69 1432.1    20.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Backward Selection Process

We will now use the backward selection process to remove any variables that are insignificant, meaning they have a p-value less than our significance threshold of 0.10. This process will remove insignificant variables one at a time starting with the one with the highest significance level. The result is our simplified model, which will be evaluated against the model assumptions and its R^2 score compared to the fully constructed model. Below are the results of each step of the backward selection process.

```
Start:  AIC=241.08
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7

       Df Sum of Sq    RSS    AIC
- x1    1       1.9 1434.0 239.18
- x6    1       4.4 1436.5 239.31
- x5    1      22.8 1454.9 240.29
<none>              1432.1 241.08
- x7    1      62.4 1494.5 242.36
- x2    1     973.1 2405.2 279.00
- x3    1    2239.9 3672.0 311.58
- x4    1    3553.9 4986.0 335.14
```

```
Step:  AIC=239.18
y ~ x2 + x3 + x4 + x5 + x6 + x7

       Df Sum of Sq    RSS    AIC
- x6    1       3.0 1437.0 237.34
- x5    1      22.9 1456.9 238.40
<none>              1434.0 239.18
- x7    1      62.5 1496.5 240.47
- x2    1     994.1 2428.1 277.73
- x3    1    2576.1 4010.1 316.36
- x4    1    4422.6 5856.7 345.53
```

```
Step:  AIC=237.34
y ~ x2 + x3 + x4 + x5 + x7

       Df Sum of Sq    RSS    AIC
- x5    1      28.8 1465.8 236.87
<none>              1437.0 237.34
- x7    1      66.0 1503.0 238.80
- x2    1    1096.4 2533.5 279.00
- x3    1    2772.8 4209.8 318.11
- x4    1    5963.0 7400.1 361.54
```

```
Step:  AIC=236.87
y ~ x2 + x3 + x4 + x7

       Df Sum of Sq    RSS    AIC
<none>              1465.8 236.87
- x7    1      56.9 1522.8 237.81
- x2    1    1384.2 2850.0 286.07
- x3    1    2747.6 4213.4 316.17
- x4    1    6129.9 7595.8 361.55
```

Below is the model summary of the final model generated after the backward selection process.

```
Call:
lm(formula = y ~ x2 + x3 + x4 + x7, data = analysis)

Residuals:
    Min      1Q  Median      3Q     Max
-17.432  -1.628   0.764   2.152   9.556

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 55.174733   2.851405  19.350  < 2e-16 ***
x2          -0.051568   0.006254  -8.245 5.34e-12 ***
x3           2.990215   0.257399  11.617  < 2e-16 ***
x4          -2.068301   0.119196 -17.352  < 2e-16 ***
x7           4.393618   2.626985   1.672   0.0988 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.512 on 72 degrees of freedom
Multiple R-squared:  0.9023,    Adjusted R-squared:  0.8968
F-statistic: 166.2 on 4 and 72 DF,  p-value: < 2.2e-16
```

After removing x1 (protein), x5 (vitamins), and x6 (weight), we end up with a R^2 score of 0.9023 and an adjusted R^2 of 0.8968. While the R^2 is slightly lower than that of the full model, the adjusted R^2 is slightly higher. Our final model is in the form of:

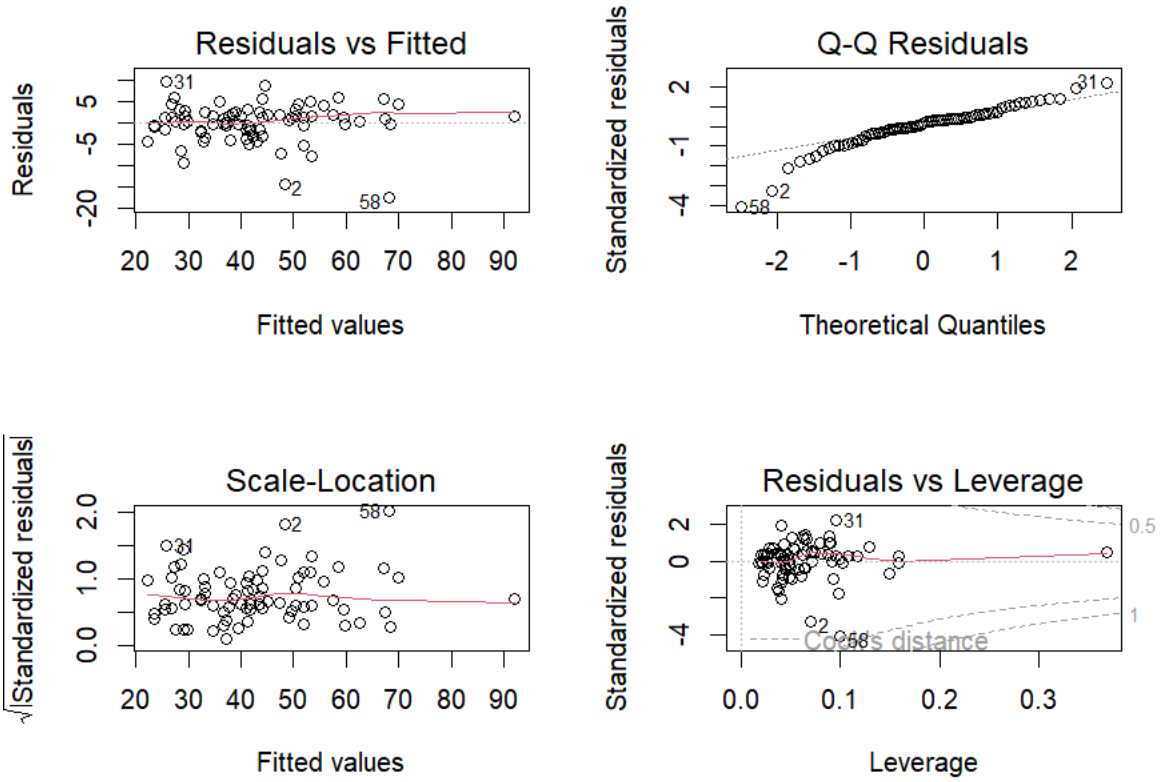
$$Y_i = \beta_0 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_7 x_{i7} + \epsilon_i$$

$$i = 1, 2, \ldots, n, where\ \epsilon_i = N(0, \sigma^2)$$

Where ϵ_i are independent and identically distributed (iid)

Model Assumption Check on the Final Model:

Before analyzing the final model generated by the backward selection process, we must verify that it still satisfies the four model assumptions like we did with the full model. For our linearity check, we observe that there is no obvious pattern in the Residuals vs. Fitted plot as indicated by the red line, which means the linearity assumption holds for this final model. We still do not need to perform a test to check for independence of errors as our data is not time-series. To check for normality of errors, we observe the results of the QQ Plot of residuals, which indicates normality of errors but with a few significant outliers. For our equal variances check, we check the plot of Standardized Residuals vs. Fitted Values for any obvious patterns. This model's plot indicates that there is no obvious pattern, reinforced by the red line on the plot. This means the final model generated by the backward selection process still satisfies all four criteria, meaning transformation is not necessary.



Conclusion

The final multilinear model and parameter estimates as derived from the backward selection process is as follows:

$$Y_i = 55.174773 - 0.051568x_2 + 2.990215x_3 - 2.068301x_4 + 4.393618x_7$$

In our final model, the variables sodium (x2), dietary fiber (x3), sugars (x4), and cups per serving (x7) appear to be the most significant in determining the rating of a brand of cereal. This simplified final model still passes the assumption checks that the full model did.

The R^2 score of our final model is 0.9023 with an adjusted R^2 of 0.8968, which means 89.68% of the variation in rating can be explained by sodium, dietary fiber, sugars, and cups per serving. Compared to the original full model, which had an R^2 of 0.9045 and an adjusted R^2 of 0.8948, the simplified model has a slightly worse R^2 but a slightly better adjusted R^2. This indicates that this simplified model still does an excellent job at determining the rating of a brand of cereal.

Per the screenshot below, the AIC and BIC for the simplified model are both lower than that of the original full model, which indicates that the simplified model has a greater goodness of fit.

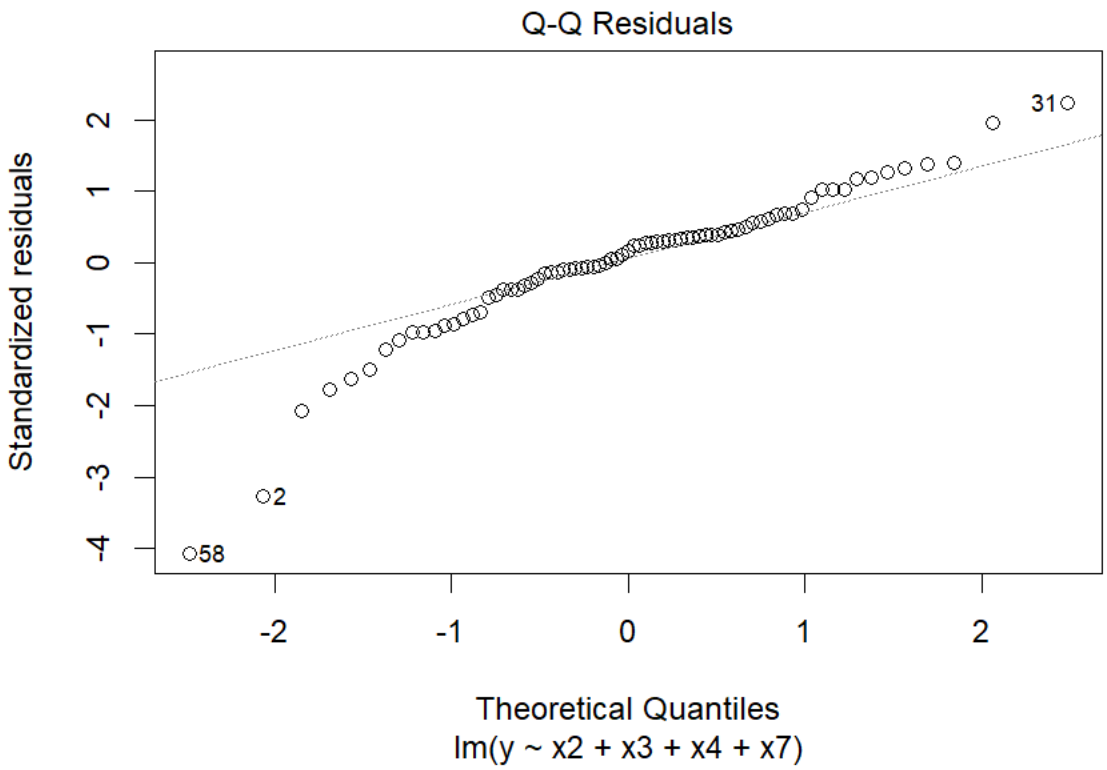
```
> # AIC and BIC comparison between the two models
> AIC(model) # 461.5946
[1] 461.5946
> BIC(model) # 482.6889
[1] 482.6889
>
> AIC(back) # 457.3878
[1] 457.3878
> BIC(back) # 471.4506
[1] 471.4506
```

To interpret the coefficients, if the other three variables are held constant…

- A one milligram increase in sodium (x2) decreases a cereal's rating by 0.051568.
- A one-gram increase in dietary fiber (x3) increases a cereal's rating by 2.990215.
- A one-gram increase in sugar (x4) decreases a cereal's rating by 2.068301.
- A one cup increase in cups per serving (x7) increases a cereal's rating by 4.393618.

Knowing this, it can be easier to quantifiably estimate how healthy a brand of cereal is using only these four characteristics, and how each of these characteristics influence a brand's health rating. Not only is the simplified model easier to interpret, but it also performs similarly well to the full, more complex model while also addressing the issue of overfitting that more complex models have. However, simpler models are not without their own tradeoffs that one must be aware of, such as higher error and bias.

Note that this analysis could be further improved using other selection methods and removing or otherwise dealing with influential points or outliers, such as the ones notated with a number in the QQ Plot of residuals below. Furthermore, other criteria such as AIC and BIC, can be used to compare the goodness of fit of each model.



Appendix (Full R Code)

```
# Andrew Gascon
# STAT 481 Project 1
# March 29, 2024

# Question: Using backward selection, can we derive a simplified multilinear
# model that still does a good job at determining a cereal brand's nutrition
# rating compared to the original full model?

# Set the working directory and load the data.
setwd("C:/Users/Owner/Downloads/STAT 481")
cereal <- read.csv(file = "CerealsRating.csv", header = TRUE)
```

```
# y: rating
# x: protein (x1), sodium (x2), fiber (x3), sugars (x4), vitamins (x5),
#    weight (x6), cups (x7)

# Select only the columns we're interested in:
analysis <- subset(cereal, select = c("rating", "protein", "sodium", "fiber",
                                      "sugars", "vitamins", "weight", "cups"))

# Convert columns into variables
y <- analysis$rating    # y = rating
x1 <- analysis$protein  # x1 = protein
x2 <- analysis$sodium   # x2 = sodium
x3 <- analysis$fiber    # x3 = fiber

x4 <- analysis$sugars   # x4 = sugars
x5 <- analysis$vitamins # x5 = vitamins
x6 <- analysis$weight   # x6 = weight
x7 <- analysis$cups     # x7 = cups

# Number of NA values in y (rating):
length(which(y == -1))

# Number of NA values in x1 (protein): 0
length(which(x1 == -1))

# Number of NA values in x2 (sodium): 0
length(which(x2 == -1))
```

```
# Number of NA values in x3 (fiber): 0
length(which(x3 == -1))

# Number of NA values in x4 (sugars): 1
length(which(x4 == -1))

# Number of NA values in x5 (vitamins): 0
length(which(x5 == -1))

# Number of NA values in x6 (weight): 0
length(which(x6 == -1))

# Number of NA values in x7 (cups): 0
length(which(x7 == -1))

# There's a missing value in sugars (x4), so we'll replace it with a 0.
analysis$sugars[x4 == -1] <- 0

# Descriptive statistics:

# Rating (y)
fivenum(y)
mean(y)
sd(y)

par(mfrow = c(2, 1))
hist(y, xlab = "Rating (y)", ylab = "Frequency",
```

```
    main = "Distribution of Rating (y)")

ybp <- boxplot(y, ylab = "Rating (y)", main = "Distribution of Rating (y)")
length(ybp$out) # Number of outliers in Rating (y): 1

# Protein (x1)
fivenum(x1)
mean(x1)
sd(x1)

hist(x1, xlab = "Protein (x1)", ylab = "Frequency",
    main = "Distribution of Protein (x1)")

x1bp <- boxplot(x1, ylab = "Protein (x1)", main = "Distribution of Protein (x1)")
length(x1bp$out) # Number of outliers in Protein (x1): 3

# Sodium (x2)
fivenum(x2)
mean(x2)
sd(x2)

hist(x2, xlab = "Sodium (x2)", ylab = "Frequency",
    main = "Distribution of Sodium (x2)")

x2bp <- boxplot(x2, ylab = "Sodium (x2)", main = "Distribution of Sodium (x2)")
length(x2bp$out) # Number of outliers in Sodium (x2): 9

# Fiber (x3)
```

```
fivenum(x3)
mean(x3)
sd(x3)

hist(x3, xlab = "Fiber (x3)", ylab = "Frequency",
    main = "Distribution of Fiber (x3)")

x3bp <- boxplot(x3, ylab = "Fiber (x3)", main = "Distribution of Fiber (x3)")
length(x3bp$out) # Number of outliers in Fiber (x3): 3

# Sugars (x4)
fivenum(x4)
mean(x4)
sd(x4)

hist(x4, xlab = "Sugars (x4)", ylab = "Frequency",
    main = "Distribution of Sugars (x4)")

x4bp <- boxplot(x4, ylab = "Sugars (x4)", main = "Distribution of Sugars (x4)")
length(x4bp$out) # Number of outliers in Sugars (x4): 0

# Vitamins (x5)
fivenum(x5)
mean(x5)
sd(x5)

hist(x5, xlab = "Vitamins (x5)", ylab = "Frequency",
    main = "Distribution of Vitamins (x5)")
```

```
x5bp <- boxplot(x5, ylab = "Vitamins (x5)", main = "Distribution of Vitamins (x5)")
length(x5bp$out) # Number of outliers in Vitamins (x5): 14

# Weight (x6)
fivenum(x6)
mean(x6)
sd(x6)

hist(x6, xlab = "Weight (x6)", ylab = "Frequency",
    main = "Distribution of Weight (x6)")

x6bp <- boxplot(x6, ylab = "Weight (x6)", main = "Distribution of Weight (x6)")
length(x6bp$out) # Number of outliers in Weight (x6): 13

# Cups (x7)
fivenum(x7)
mean(x7)
sd(x7)

hist(x7, xlab = "Cups (x7)", ylab = "Frequency",
    main = "Distribution of Cups (x7)")

x7bp <- boxplot(x7, ylab = "Cups (x7)", main = "Distribution of Cups (x7)")
length(x7bp$out) # Number of outliers in Cups (x7): 1

# Full model and model check.
model <- lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7, data = analysis)
```

```
par(mfrow = c(1, 1))
plot(model)

# Multicolinearity check
library(car)
vif(model)

# Variable selection:
summary(model)
anova(model)

# Backward selection:
back <- step(model, direction = "backward", trace = 0)
summary(back)
anova(back)

# Model check for the simplified model.
par(mfrow = c(1, 1))
plot(back, ask = FALSE)

# AIC and BIC comparison between the two models
AIC(model) # 461.5946
BIC(model) # 482.6889

AIC(back) # 457.3878
BIC(back) # 471.4506
```