Multiple Linear Regression: Cereal Ratings

Ellen Chancey

December 10, 2017

I. Introduction

A. Study Design

A dataset containing nutritional information and consumer ratings of 77 varieties of cereal is obtained and is to be explored for a potential linear regression model. The outcome variable for this model, Y, is the consumer rating of the cereal, with nutritional data making up the predictor variables. These include the amount of calories, protein, fat, sodium, fiber, carbohydrates, sugar, and potatssium in a reccomended serving of the cereal. Data was obtained from kaggle.

B. Aims

The purpose of this investigation is to determine a model that can estimate consumer ratings based on the nutritional content of the cereal in a parsimonious manner. The goal of the model is to indicate which nutritional data points may best indicate a highly regarded cereal.

C. Statistical Model

A multiple linear regression model is considered. Let

$Y_i =$ the consumer rating for the i^{th} cereal

$X_{i1} =$ the number of calories for the i^{th} cereal,

$X_{i2} =$ the amount of protein (in grams) of the i^{th} cereal,

$X_{i3} =$ the amount of fat (in grams) of the i^{th} cereal,

$X_{i4} =$ the amount of sodium (in milligrams) of the i^{th} cereal,

$X_{i5} =$ the amount of fiber (in grams) of the i^{th} cereal,

$X_{i6} =$ the amount of carbohydrates (in grams) of the i^{th} cereal,

$X_{i7} =$ the amount of sugar (in grams) of the i^{th} cereal,

$X_{i8} =$ the amount of potassium (in milligrams) of the i^{th} cereal.

The **initial model** is given by

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \beta_7 X_{i7} + \beta_8 X_{i8} + \varepsilon_i$$

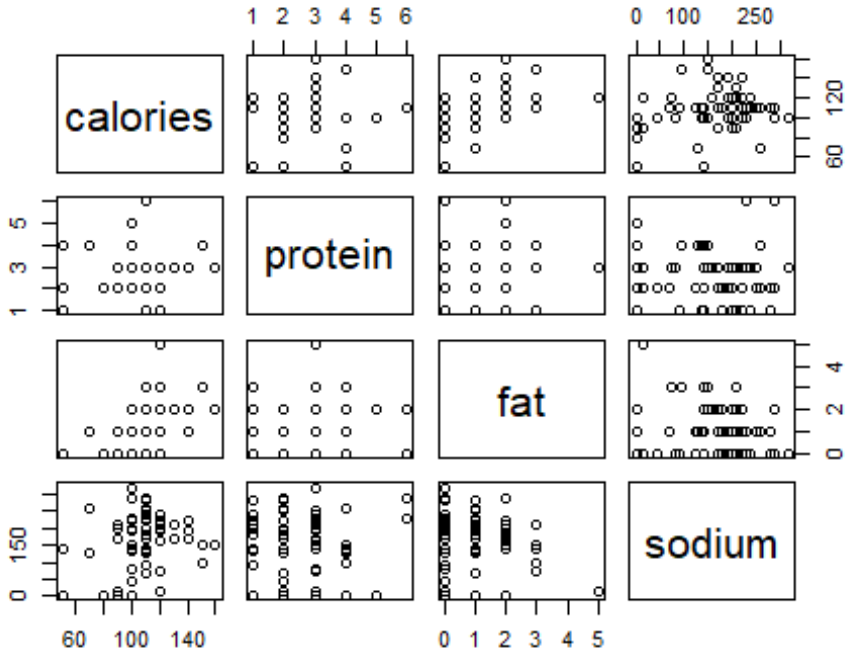
where $\varepsilon_i \sim iidN(0, \sigma^2), i = 1,2,..,77$, and $\beta_0, \beta_1, \ldots, \beta_8$, and σ^2 are the unknown model parameters.

II. Preliminary Analysis

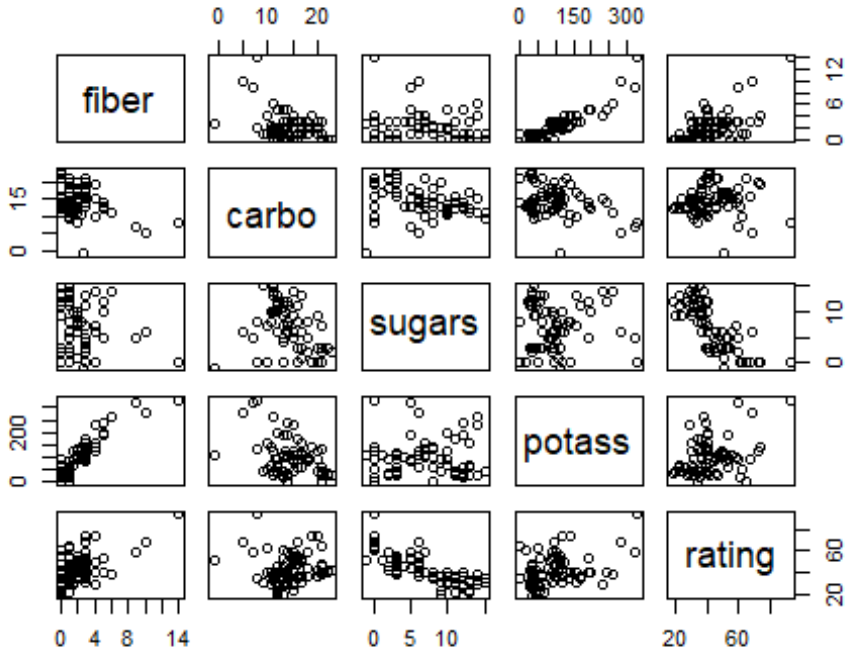
A. Bivariate Associations

A pair of scatterplot matrices displays the relationship of all variables.

```
pairs(C2[1:4])
```




```
pairs(C2[5:9])
```



The following table provides the correlation coefficient of all associations in the dataset. Rating is most closely associated with sugar and calorie content.

```
cor(C2)
```

```
##              calories     protein           fat       sodium       fiber
## calories   1.00000000  0.01906607  0.498609814  0.300649227 -0.29341275
## protein    0.01906607  1.00000000  0.208430990 -0.054674348  0.50033004
## fat        0.49860981  0.20843099  1.000000000 -0.005407464  0.01671924
## sodium     0.30064923 -0.05467435 -0.005407464  1.000000000 -0.07067501
## fiber     -0.29341275  0.50033004  0.016719237 -0.070675009  1.00000000
## carbo      0.25068091 -0.13086365 -0.318043492  0.355983473 -0.35608274
## sugars     0.56234029 -0.32914178  0.270819175  0.101451381 -0.14120539
## potass    -0.06660886  0.54940740  0.193278602 -0.032603467  0.90337367
## rating    -0.68937603  0.47061846 -0.409283660 -0.401295204  0.58416042
##                carbo      sugars      potass      rating
## calories   0.25068091  0.56234029 -0.06660886 -0.68937603
## protein   -0.13086365 -0.32914178  0.54940740  0.47061846
## fat       -0.31804349  0.27081918  0.19327860 -0.40928366
## sodium     0.35598347  0.10145138 -0.03260347 -0.40129520
## fiber     -0.35608274 -0.14120539  0.90337367  0.58416042
## carbo      1.00000000 -0.33166538 -0.34968522  0.05205466
## sugars    -0.33166538  1.00000000  0.02169581 -0.75967466
## potass    -0.34968522  0.02169581  1.00000000  0.38016537
## rating     0.05205466 -0.75967466  0.38016537  1.00000000
```

B. Screening of Covariates and Verification of Assumptions

The following method was used to select the final predictor variables that offer a parsimonious model. First, automatic variables selection was used to develop candidate models. Second, three criterion were considered for each model: C_p, BIC, and adjusted R^2. Partial residual plots, residual-versus-fitted plots, and measures of influence are investigated. One observation was identified as an influential outlier and was excluded from the final model. With that exclusion, there were no problems with the linearity, constant variance, independence, or normality of model residuals.

Through this process, the following variables were removed from the model: calories, protein, carbohydrates, and potassium.

C. Final Model

The **final model** is given by

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i$$

where $Y_i =$ the consumer rating for the i^{th} cereal

$X_{i1} =$ the amount of fat (in grams) of the i^{th} cereal,

$X_{i2} =$ the amount of sodium (in milligrams) of the i^{th} cereal,

$X_{i3} =$ the amount of fiber (in grams) of the i^{th} cereal,

$X_{i4} =$ the amount of sugar (in grams) of the i^{th} cereal.

where $\varepsilon_i \sim iidN(0, \sigma^2), i = 1,2,\ldots,77$, and $\beta_0, \beta_1, \ldots, \beta_4$, and σ^2 are the unknown model parameters.

III. Statistical Analysis

According to the fitted model, the best variables to consider when attempting to predict consumer rating are fat, sodium, fiber, and sugar. This model accounts for % of variability in the data.

```
model5sum

##
## Call:
## lm(formula = rating ~ fat + sodium + fiber + sugars, data = C2,
##     subset = -58)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3394 -1.3669 -0.2298  1.1915  7.3055
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 62.535230   0.818157   76.43   <2e-16 ***
## fat         -3.325458   0.287630  -11.56   <2e-16 ***
## sodium      -0.055642   0.003358  -16.57   <2e-16 ***
## fiber        2.832353   0.116396   24.33   <2e-16 ***
## sugars      -1.953304   0.066835  -29.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.386 on 71 degrees of freedom
## Multiple R-squared:  0.9729, Adjusted R-squared:  0.9714
## F-statistic:   638 on 4 and 71 DF,  p-value: < 2.2e-16
```

Correlation between ratings is highest with sugar content with -0.7596747, followed by fiber with 0.5841604, fat with -0.4092837, and -0.4012952 for sodium.

```
cor(C3)

##                  fat       sodium       fiber     sugars     rating
## fat     1.000000000 -0.005407464  0.01671924  0.2708192 -0.4092837
## sodium -0.005407464  1.000000000 -0.07067501  0.1014514 -0.4012952
## fiber   0.016719237 -0.070675009  1.00000000 -0.1412054  0.5841604
## sugars  0.270819175  0.101451381 -0.14120539  1.0000000 -0.7596747
## rating -0.409283660 -0.401295204  0.58416042 -0.7596747  1.0000000
```

IV. Summary of Findings

Based on these findings, the best nutritional components of cereal that can predict consumer ratings includes sugars, fiber, fat, and sodium. Higher sugar, fat, and sodium contents are associated with a lower rating, while cereals high in fiber are associated with a higher rating.

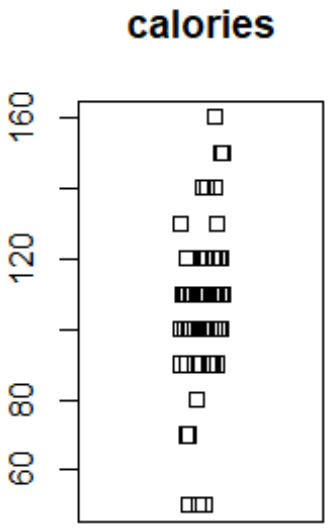
V. Appendix

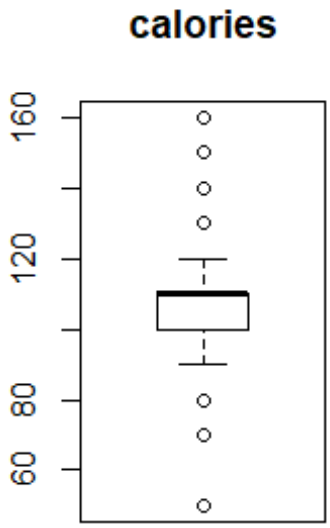
A. Diagnostics for Predictors

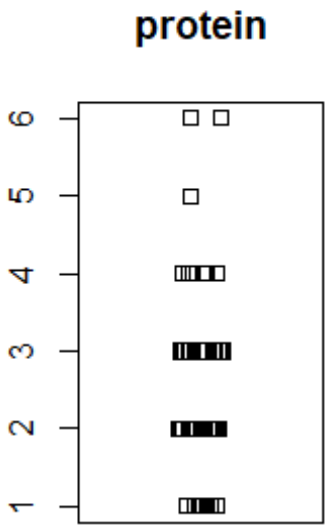
To evaluate the potential predictors under consideration boxplots and stripcharts of each variable are considered.

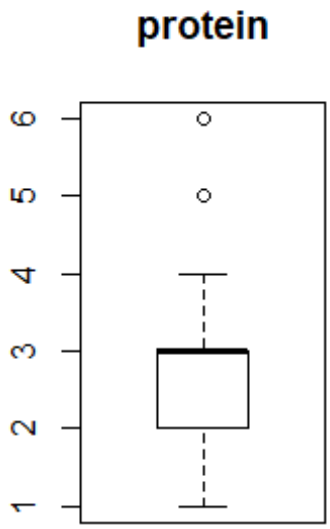
Comments
* Many variables have clustered values including calories, protein, fat, and fiber
* Fiber and potassium are skewed
* No variables are eliminated based on this information.

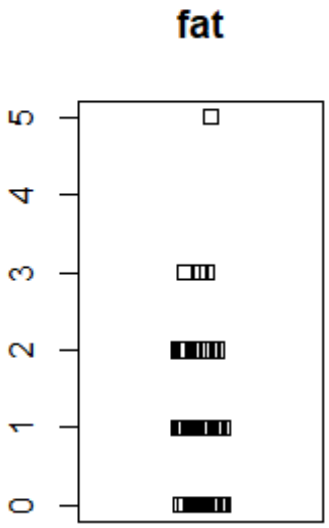
```
# for loop for boxplots and strip charts
for (i in 1:8){
  par(mfrow=c(1,2))
  stripchart(C2[,i], main = names(C2)[i],vertical = T, method = "jitter")
  boxplot(C2[,i], main = names(C2)[i])
  par(mfrow=c(1,1))
}
```

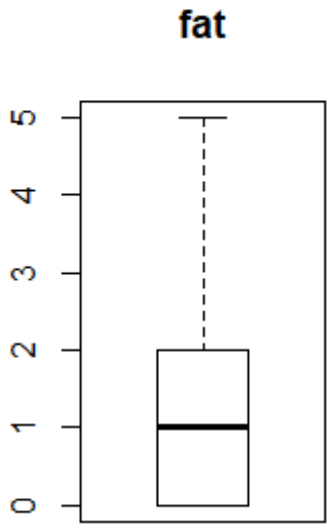

calories
160
120
80
60

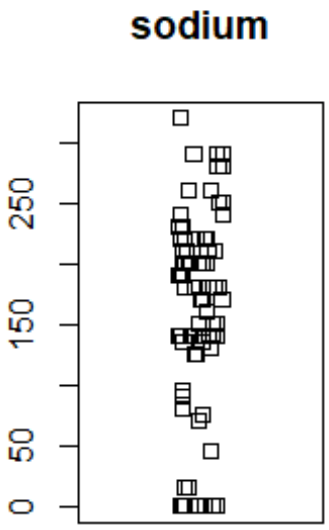

calories
160
120
80
60

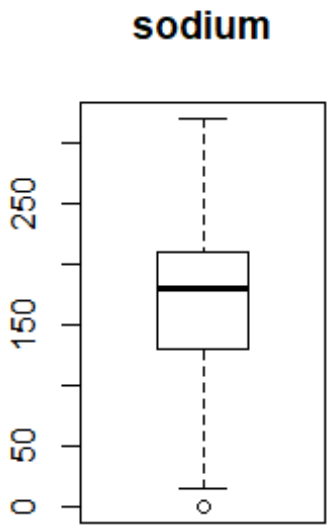

protein
6
5
4
3
2
1

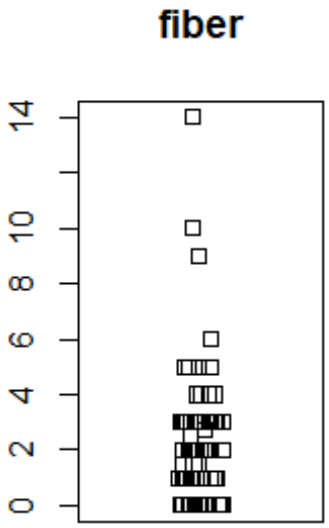

protein
6
5
4
3
2
1

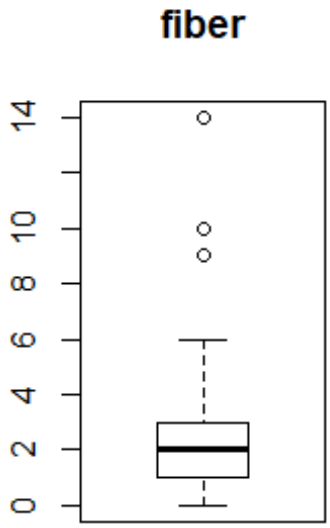

fat
0
1
2
3
4
5

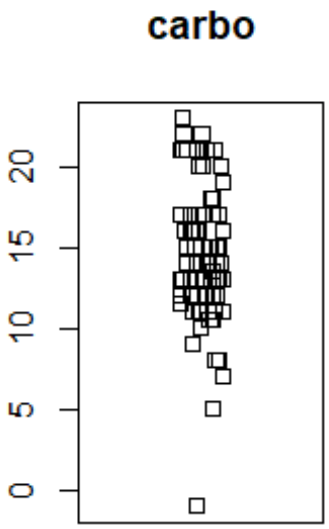

fat
0
1
2
3
4
5

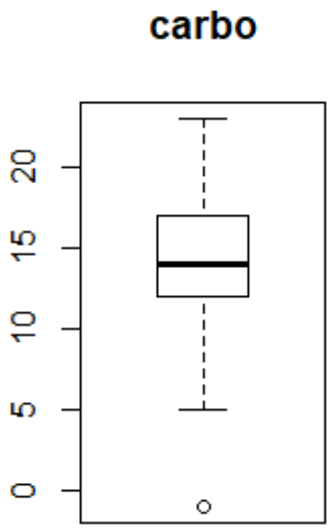

sodium
0
50
150
250

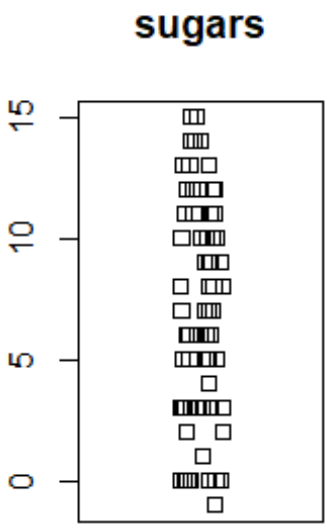

sodium
0
50
150
250

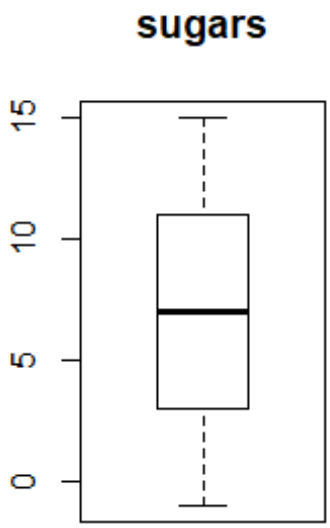

fiber
0
2
4
6
8
10
14

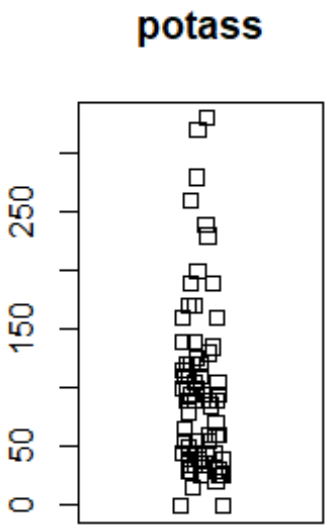

fiber
0
2
4
6
8
10
14

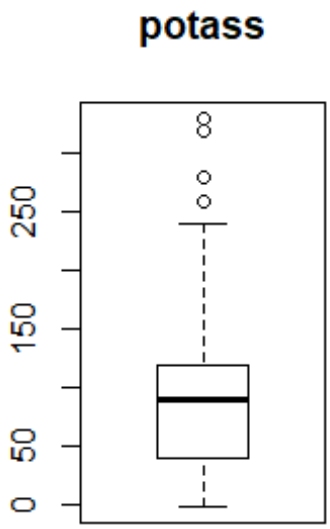

carbo
0
5
10
15
20


carbo
0
5
10
15
20


sugars
0
5
10
15


sugars
0
5
10
15


potass
0
50
150
250

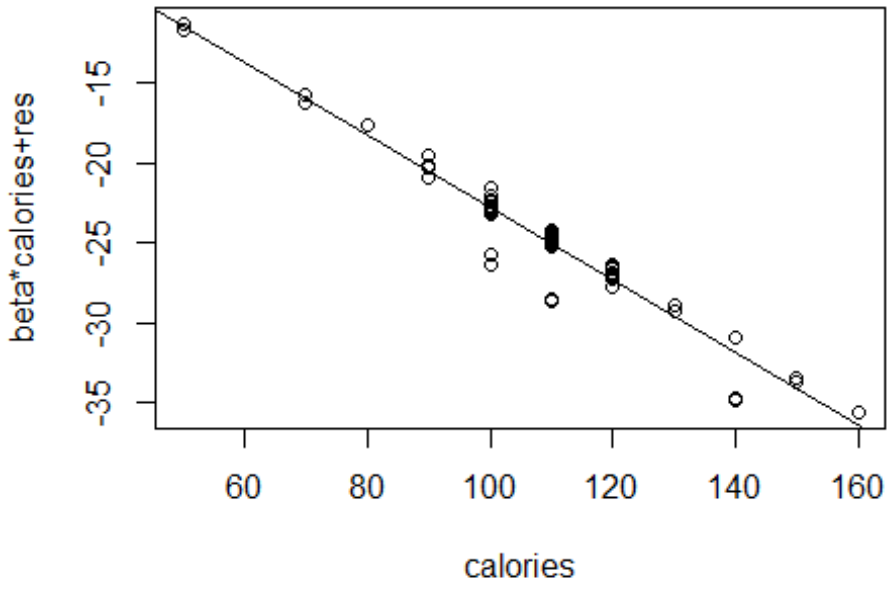

potass
0
50
150
250

B. Screening of Predictors

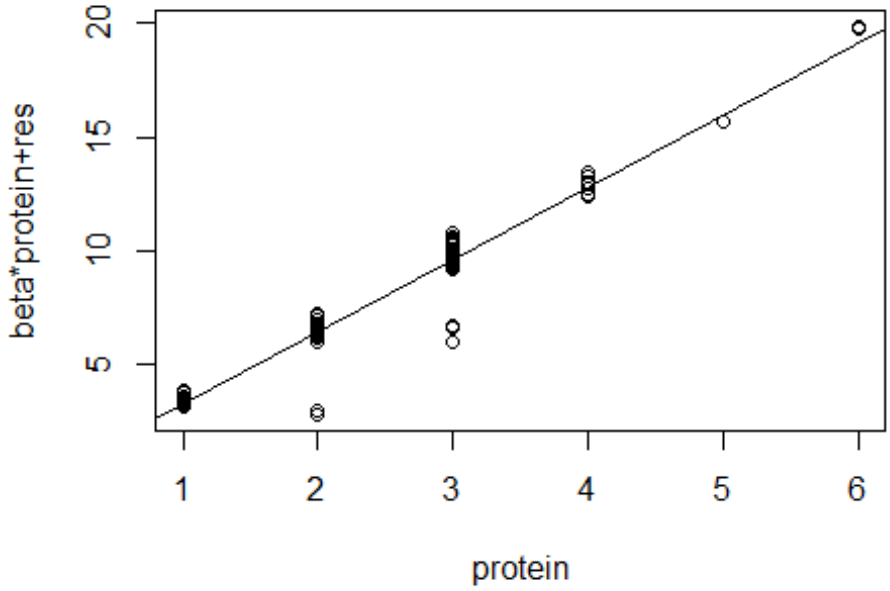
1. Added variable plots

The following figures aim to determine if predictor variables conveys helpful information for predicting consumer ratings. Based on these plots, all variables appear to have an impact on ratings. No predictor variables are ruled out at this point.

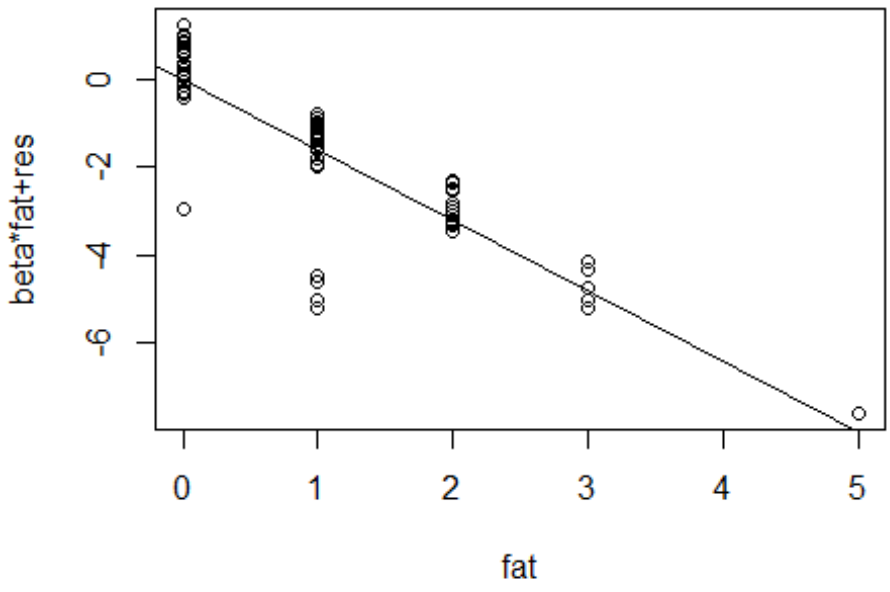
```
prplot(model,1)
```



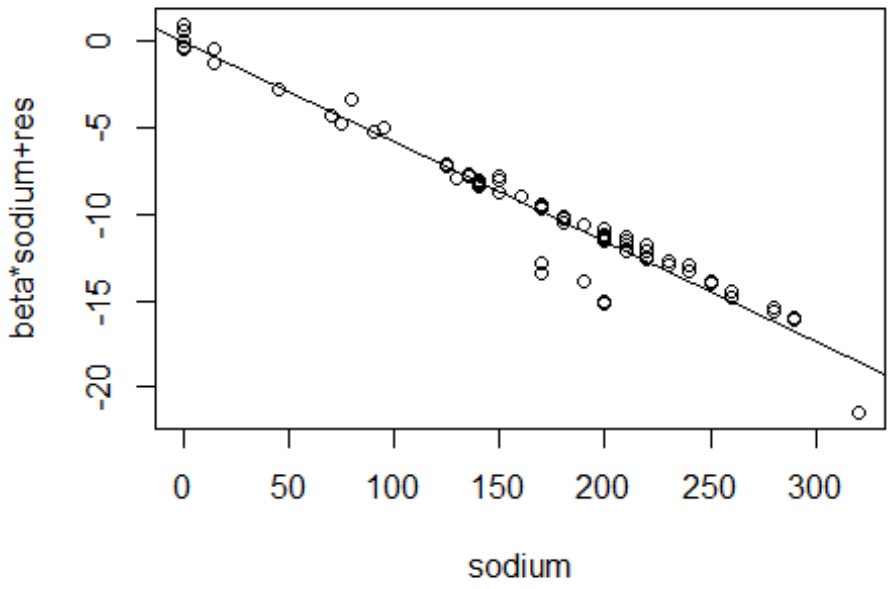

```
prplot(model,2)
```



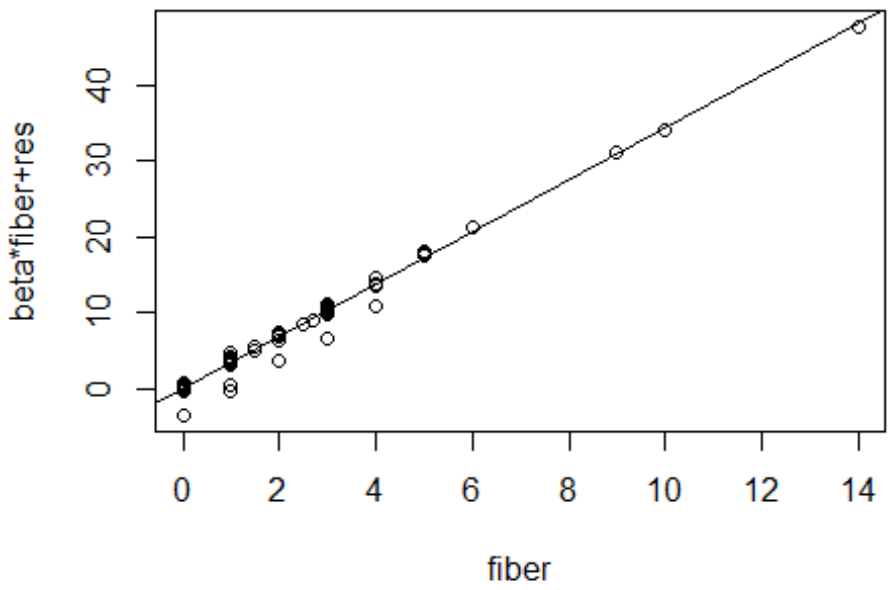

```
prplot(model,3)
```



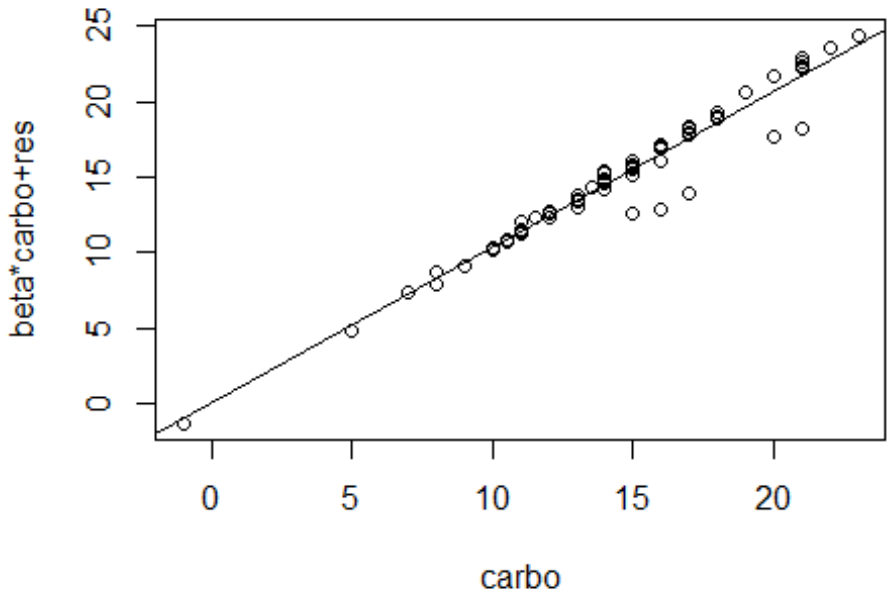

```
prplot(model,4)
```



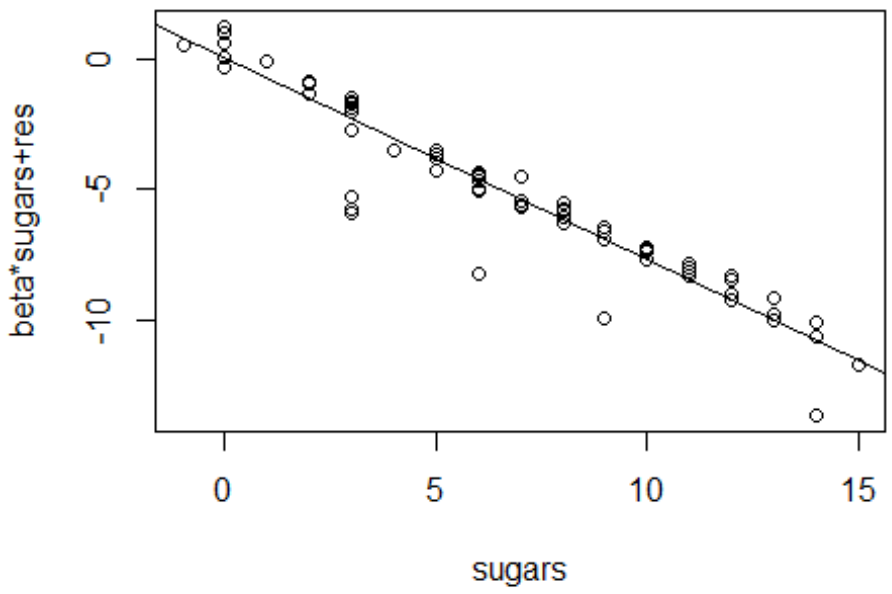

```
prplot(model,5)
```



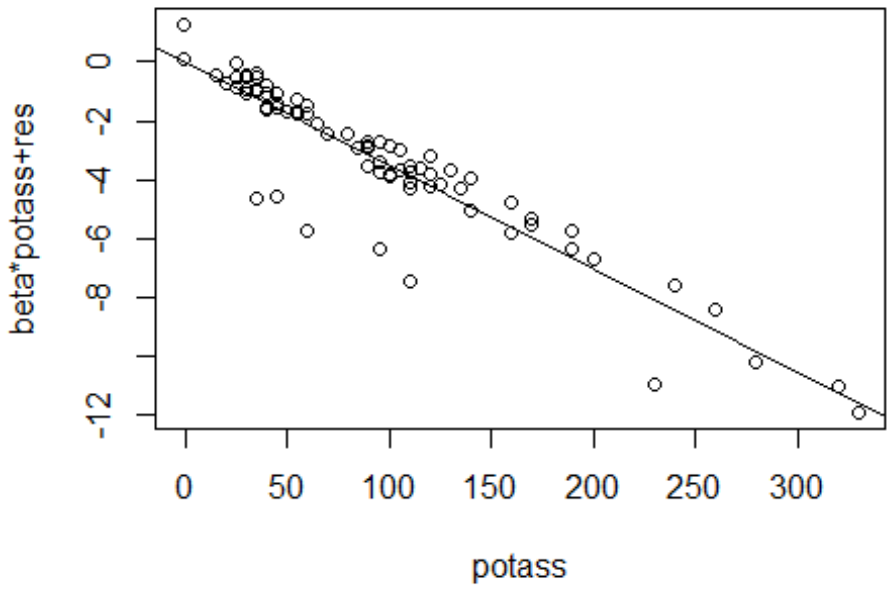

```
prplot(model,6)
```




```
prplot(model,7)
```



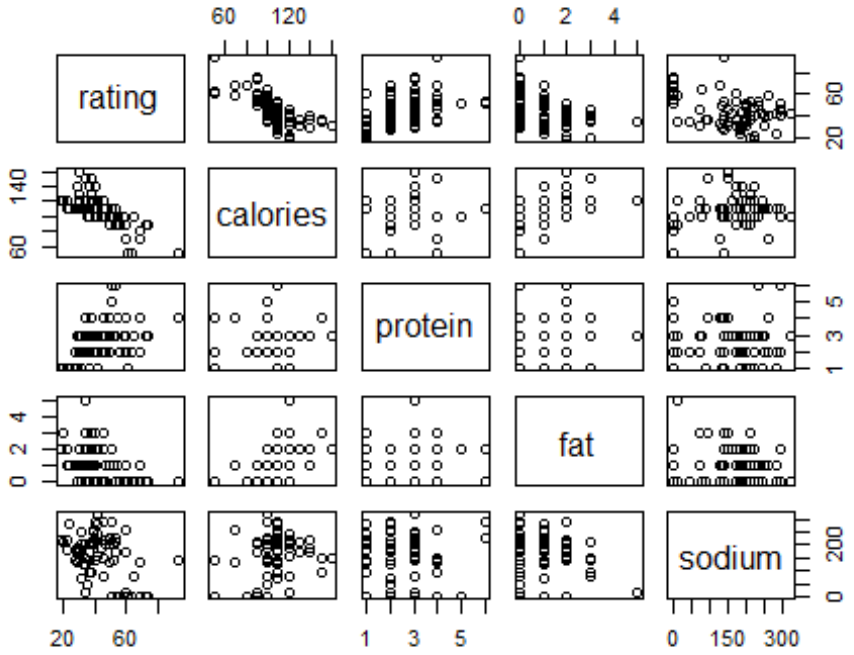

```
prplot(model,8)
```



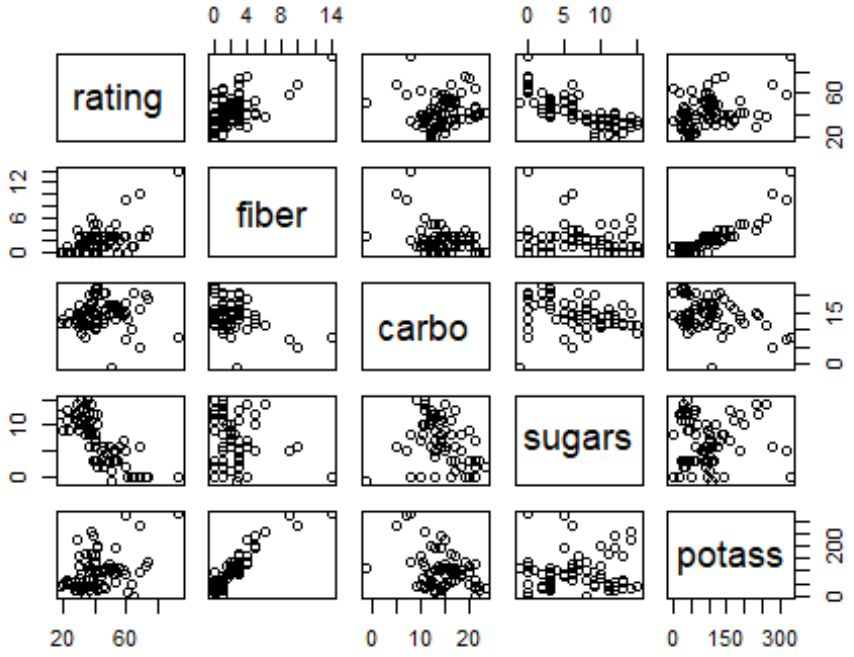
2. Evaluation of Multicolinearity

The following pair of scatterplots displays all potential predictor variables and is considered to determine if any variables are candidates for multicolinearity. These plots indicate that potassium and fiber have a linear relationship. These will be investigated further.

```
pairs(rating ~ calories + protein + fat + sodium, data = C2)
```




```
pairs(rating ~ fiber + carbo + sugars + potass, data = C2)
```



Pearson's correlation coefficient is also considered, and interpretation of these values is easier than reading the many scatterplots produced above. These values identify additional relationships. Relationships between potassium and fiber, potassium and protein, and sugar and calories are noted.

No variables are eliminated at this point. Formal measures of mutlicolinearity will be considered when the final model is selected.

```
cor(C2)
```

```
##             calories     protein          fat       sodium       fiber
## calories  1.00000000  0.01906607  0.498609814  0.300649227 -0.29341275
## protein   0.01906607  1.00000000  0.208430990 -0.054674348  0.50033004
## fat       0.49860981  0.20843099  1.000000000 -0.005407464  0.01671924
## sodium    0.30064923 -0.05467435 -0.005407464  1.000000000 -0.07067501
## fiber    -0.29341275  0.50033004  0.016719237 -0.070675009  1.00000000
## carbo     0.25068091 -0.13086365 -0.318043492  0.355983473 -0.35608274
## sugars    0.56234029 -0.32914178  0.270819175  0.101451381 -0.14120539
## potass   -0.06660886  0.54940740  0.193278602 -0.032603467  0.90337367
## rating   -0.68937603  0.47061846 -0.409283660 -0.401295204  0.58416042
##                carbo      sugars      potass      rating
## calories  0.25068091  0.56234029 -0.06660886 -0.68937603
## protein  -0.13086365 -0.32914178  0.54940740  0.47061846
## fat      -0.31804349  0.27081918  0.19327860 -0.40928366
## sodium    0.35598347  0.10145138 -0.03260347 -0.40129520
## fiber    -0.35608274 -0.14120539  0.90337367  0.58416042
## carbo     1.00000000 -0.33166538 -0.34968522  0.05205466
## sugars   -0.33166538  1.00000000  0.02169581 -0.75967466
## potass   -0.34968522  0.02169581  1.00000000  0.38016537
## rating    0.05205466 -0.75967466  0.38016537  1.00000000
```

3. Automatic variable selection methods

Automatic variable selection is employed here to identify the best parsimonious models that should be considered.

```
model2 <- regsubsets(rating ~ calories + protein + fat + sodium + fiber +
carbo + sugars + potass, data = C2)
modelsum <- summary(model2)
modelsum
```

```
## Subset selection object
## Call: regsubsets.formula(rating ~ calories + protein + fat + sodium +
##     fiber + carbo + sugars + potass, data = C2)
## 8 Variables  (and intercept)
##          Forced in Forced out
## calories     FALSE      FALSE
## protein      FALSE      FALSE
## fat          FALSE      FALSE
## sodium       FALSE      FALSE
## fiber        FALSE      FALSE
```

```
## carbo         FALSE      FALSE
## sugars        FALSE      FALSE
## potass        FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          calories protein fat sodium fiber carbo sugars potass
## 1  ( 1 ) " "      " "     " " " "    " "   " "   "*"    " "
## 2  ( 1 ) " "      " "     " " " "    "*"   " "   "*"    " "
## 3  ( 1 ) " "      " "     " " "*"    "*"   " "   "*"    " "
## 4  ( 1 ) " "      " "     "*" "*"    "*"   " "   "*"    " "
## 5  ( 1 ) "*"      "*"     " " "*"    "*"   "*"   " "    " "
## 6  ( 1 ) "*"      "*"     " " "*"    "*"   "*"   "*"    " "
## 7  ( 1 ) "*"      "*"     "*" "*"    "*"   "*"   "*"    " "
## 8  ( 1 ) "*"      "*"     "*" "*"    "*"   "*"   "*"    "*"
```

Three criterion are considered for the models developed by the automatic variable selection function. These include C_p, BIC and adjusted R^2. In this instance, none of the C_p values are good, the adjusted R^2 gets sufficiently high begining with the fourth model and does not lower, even though adjusted R^2 punishes the addition of variables. Later models have good BIC values as well.

Given that the goal of this analysis is to determine a parsimonious model, model four is selected for moving forward. This model has an adjusted R^2 value of 0.9582936 and a BIC of -227.08092. Model four is the model with the fewest predictor variables that obtains sufficient, although not the best, criterion values.

```
modelsum$cp # want it close to p (number of variables)

## [1] 5305.67951 2356.28340 1222.47280  435.53372  271.72709  151.90943
## [7]   55.29423    9.00000

modelsum$adjr2 # want it high

## [1] 0.5714670 0.8039988 0.8942861 0.9582936 0.9716607 0.9816546 0.9899289
## [8] 0.9940246

modelsum$bic # want it low

## [1]  -57.58111 -114.50391 -158.74629 -227.08092 -253.56715 -283.80066
## [7] -326.74203 -363.71764
```

The recommended model contains predictor variables fat, sodium, fiber, and sugar.

```
# model four
model4 <- lm(rating ~ fat + sodium + fiber + sugars, data = C2)
model4sum <- summary(model4)
model4sum

##
## Call:
## lm(formula = rating ~ fat + sodium + fiber + sugars, data = C2)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.8547  -1.4005  -0.3453   1.5093   7.8050
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 61.364116   0.952838   64.40   <2e-16 ***
## fat         -3.618804   0.340377  -10.63   <2e-16 ***
## sodium      -0.051725   0.003954  -13.08   <2e-16 ***
## fiber        2.849014   0.139914   20.36   <2e-16 ***
## sugars      -1.864221   0.078177  -23.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.869 on 72 degrees of freedom
## Multiple R-squared:  0.9605, Adjusted R-squared:  0.9583
## F-statistic: 437.6 on 4 and 72 DF,  p-value: < 2.2e-16
```

4. Formal testing of multicolinearity

It is important at this stage to consider if multicolinearity exists in the selected model, because multicolinearity existed between some variables in the full model.
The variance inflation factor of each variable is considered. Values above ten indicate the existence of multicolinearity. The VIF values for this model are very low and indicate that multicolinearity is not present.

```
vif(model4)
```

```
##      fat   sodium    fiber   sugars
## 1.083800 1.014702 1.026897 1.115088
```

5. Preliminary model

The preliminary model is shown below.

```
model4sum
```

```
##
## Call:
## lm(formula = rating ~ fat + sodium + fiber + sugars, data = C2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.8547  -1.4005  -0.3453   1.5093   7.8050
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 61.364116   0.952838   64.40   <2e-16 ***
## fat         -3.618804   0.340377  -10.63   <2e-16 ***
## sodium      -0.051725   0.003954  -13.08   <2e-16 ***
```

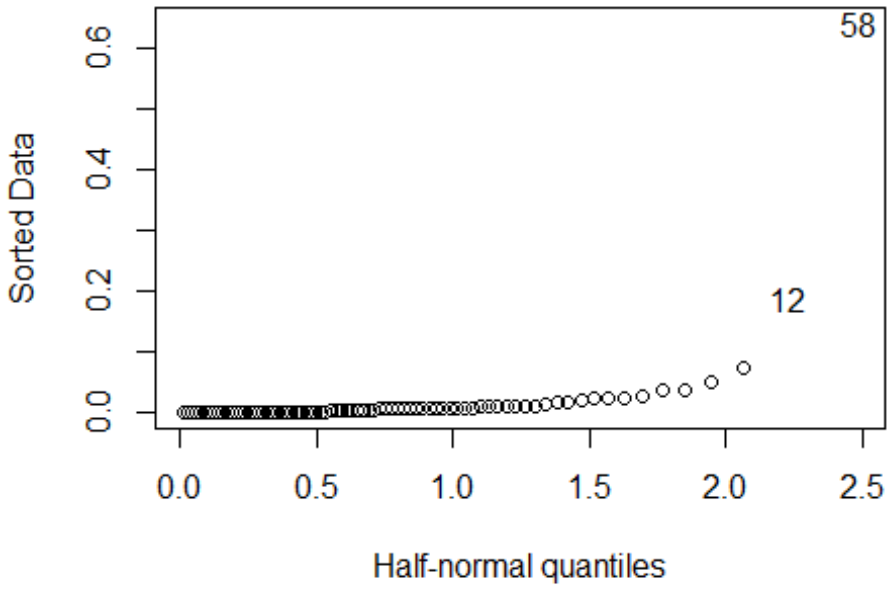
```
## fiber         2.849014   0.139914   20.36   <2e-16 ***
## sugars       -1.864221   0.078177  -23.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.869 on 72 degrees of freedom
## Multiple R-squared:  0.9605, Adjusted R-squared:  0.9583
## F-statistic: 437.6 on 4 and 72 DF,  p-value: < 2.2e-16
```

C. Residual Diagnostics

1. Influence

Cook's Distance is considered here to identify any observations that have an inflated influence on the model. The figure below graphically represents each observation's Cook's Distance. Based on this, observation 58 should be further evaluated and considered for exclusion.

```
# cook's distance
halfnorm(cooks.distance(model4))
```



Consider the details of observation 58. A negative sugar value may be a result of data entry error. In addition, the sodium value for this observation is one of the smallest.

```
C3[58,]
```

```
##    fat sodium fiber sugars   rating
## 58   2      0   2.7     -1 50.82839
```

In order to determine if this observation has an undue influence on the model, the model is fitted without observation 58. Doing so improves both the adjusted R^2 value and the F statistic. As a result, this observation will be exlcuded from analysis.

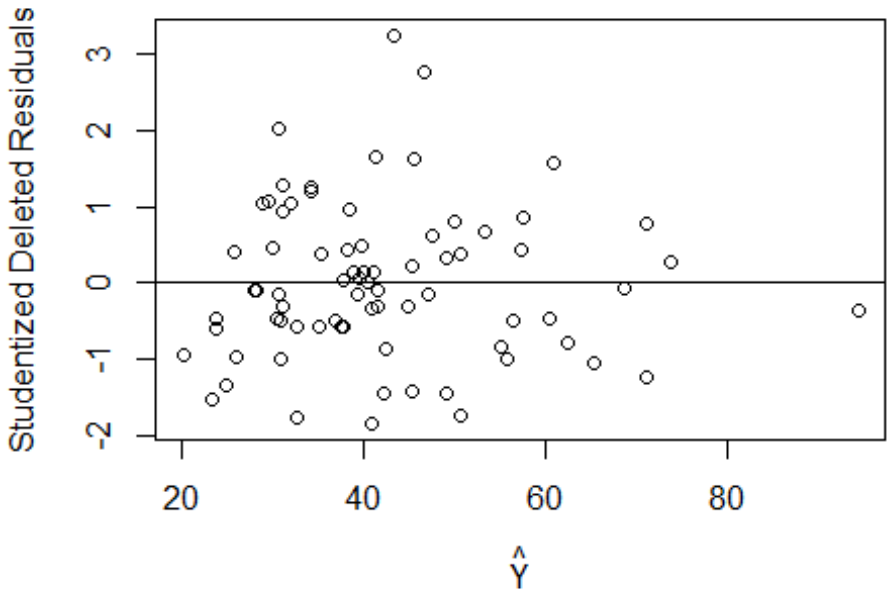
```
# fit the model without that obs
summary(lm(rating ~ fat + sodium + fiber + sugars, data = C2, subset = -58))

##
## Call:
## lm(formula = rating ~ fat + sodium + fiber + sugars, data = C2,
##     subset = -58)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3394 -1.3669 -0.2298  1.1915  7.3055
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 62.535230   0.818157   76.43   <2e-16 ***
## fat         -3.325458   0.287630  -11.56   <2e-16 ***
## sodium      -0.055642   0.003358  -16.57   <2e-16 ***
## fiber        2.832353   0.116396   24.33   <2e-16 ***
## sugars      -1.953304   0.066835  -29.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.386 on 71 degrees of freedom
## Multiple R-squared:  0.9729, Adjusted R-squared:  0.9714
## F-statistic:   638 on 4 and 71 DF,  p-value: < 2.2e-16

# establish new model
model5 <- lm(rating ~ fat + sodium + fiber + sugars, data = C2, subset = -58)
```

2. Normality and constant variance

The normality of residuals is evaluated for normality and constancy in the following figures.

Looking at the studentized deleted residuals, there is dispersion around zero without a distinct trend. This figure suggests a linear relationship and normal residuals.

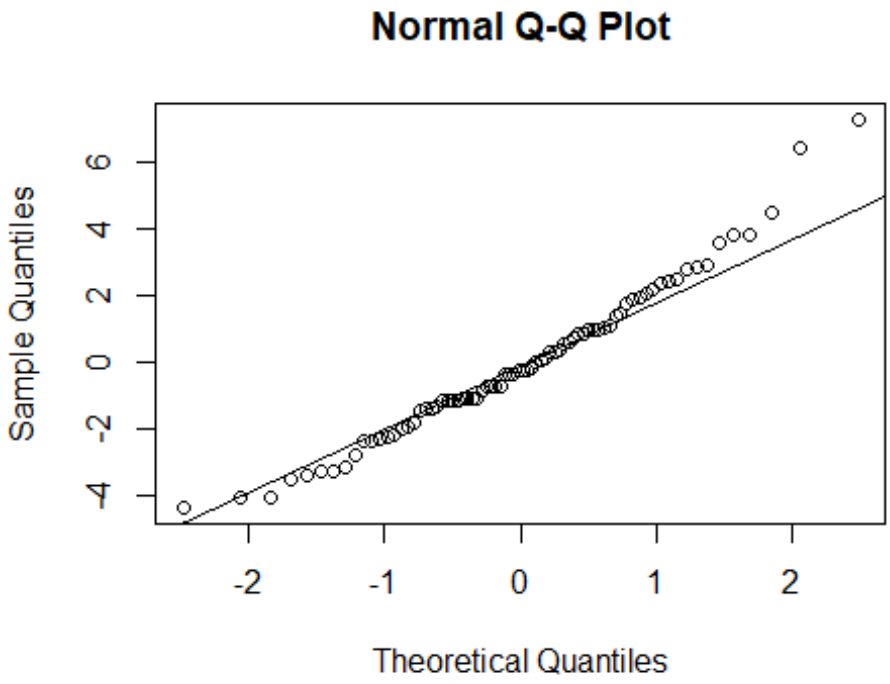
```
# studentized deleted residuals
plot(rstandard(model5)~predict(model5), xlab = expression(hat(Y)), ylab =
"Studentized Deleted Residuals")
abline(h=0)
```




```
# consistent
```

The QQ plot for this model indicates that normality of residuals is upheld.

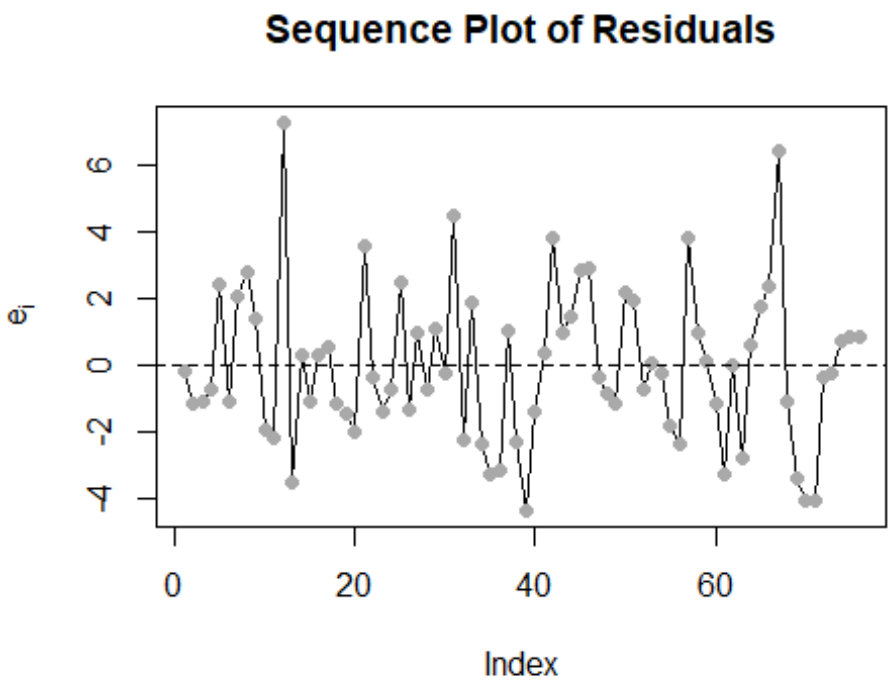
```
# qqplot
qqnorm(residuals(model5))
qqline(residuals(model5))
```



3. Independence

To evaluate the independence of residuals, a sequence plot is considered. The sequence plot indicates that there is no trend in residuals for observation. Based on this, residuals are determined to be independent.

```
# sequence plot
plot(residuals(model5),type="l",ylab=expression(e[i]),main="Sequence Plot of
Residuals")
points(residuals(model5),pch=16,col="darkgray")
abline(0,0,lty=2)
```



D. Full Script

```
# cereal final project

# https://www.kaggle.com/crawford/80-cereals/data

setwd("C:/Users/EC052367/Documents/MS-ASA/STAT 840 Linear Regression/Final
Projects/cereal")

library(leaps)
library(faraway)


######################## initial data evaluation

cereal <- read.csv("cereal.csv")
summary(cereal)

xlist <- c(4:16)
C1 <- cereal[,xlist]
attach(C1)

# for loop for boxplots and strip charts
for (i in 1:13){
  par(mfrow=c(1,2))
  stripchart(C1[,i], main = names(C1)[i],vertical = T, method = "jitter")
```

```r
  boxplot(C1[,i], main = names(C1)[i])
  par(mfrow=c(1,1))
  }
# exclude weight, shelf, vitamins, cups

######################## final dataset selection

C2xlist <- c(4,5,6,7,8,9,10,11,16)
C2 <- cereal[,C2xlist]

pairs(C2[1:6])
pairs(C2[7:13])

# rating (y)
# calories
# sugar

cor(C2)
# sugar
# calories
# fiber
# protein

# for loop for boxplots and strip charts
for (i in 1:8){
  par(mfrow=c(1,2))
  stripchart(C2[,i], main = names(C2)[i],vertical = T, method = "jitter")
  boxplot(C2[,i], main = names(C2)[i])
  par(mfrow=c(1,1))
}

# check for multicolinearity on full model

model <- lm(rating ~ calories + protein + fat + sodium + fiber + carbo +
sugars + potass, data = C2)
summary(model)

# evaluating colinearity
pairs(rating ~ calories + protein + fat + sodium, data = C2)
pairs(rating ~ fiber + carbo + sugars + potass, data = C2)
cor(C2)
# there is not multicolinearity here, largest is sugar with calories



######################## automoated variable selection
# regsubsets from leaps package
model2 <- regsubsets(rating ~ calories + protein + fat + sodium + fiber +
carbo + sugars + potass, data = C2)
modelsum <- summary(model2)
```

```r
modelsum

modelsum$cp # want it close to p (number of variables)
modelsum$adjr2 # want it high
modelsum$bic # want it low

# none of the Cp values are good, model four gets a lot of adjusted R2, and
has a low, but not the lowest model
# keeping a parsimonious model in mind
### model four is selected


####################### set up selected model

C3xlist <- c(3,4,5,7,9)
C3 <- C2[,C3xlist]

model4 <- lm(rating ~ fat + sodium + fiber + sugars, data = C2)
model4sum <- summary(model4)
model4sum


####################### evaluate selected model

# added variable plots
prplot(model,1)
prplot(model,2)
prplot(model,3)
prplot(model,4)
prplot(model,5)
prplot(model,6)
prplot(model,7)
prplot(model,8)

############## multicolinearity
pairs(rating ~ fat + sodium + fiber + sugars, data = C2)
cor(C3)

## Variance Inflation Factor  (faraway package)
# none should be above 10
vif(model4)
# all are very low

############## outliars
# cook's distance
halfnorm(cooks.distance(model4))
C3[58,]
# obs 58 may be an outlier
```

```r
# fit the model without that obs
summary(lm(rating ~ fat + sodium + fiber + sugars, data = C2, subset = -58))
# removing this obs does improve the model
# higher f stat, adjusted r2

model5 <- lm(rating ~ fat + sodium + fiber + sugars, data = C2, subset = -58)
model5sum <- summary(model5)

############### normality

# studentized deleted residuals
plot(rstandard(model5)~predict(model5), xlab = expression(hat(Y)), ylab =
"Studentized Deleted Residuals")
abline(h=0)
# consistent

# qqplot
qqnorm(residuals(model5))
qqline(residuals(model5))

############### independence

# sequence plot
plot(residuals(model5),type="l",ylab=expression(e[i]),main="Sequence Plot of
Residuals")
points(residuals(model5),pch=16,col="darkgray")
abline(0,0,lty=2)
summary(lm(residuals(model5)[-1]~-1+residuals(model5)[-47]))
# looks good


######################### final model

summary(model5)
anova(model5)
confint(model5)

pairs(rating ~ fat + sodium + fiber + sugars, subset = -58, data = C2)
cor(C3)

## R version 3.4.1 (2017-06-30)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 15063)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
```

```
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] faraway_1.0.7 leaps_3.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.13    lattice_0.20-35 digest_0.6.12   rprojroot_1.2
##  [5] MASS_7.3-47     grid_3.4.1      nlme_3.1-131    backports_1.1.0
##  [9] magrittr_1.5    evaluate_0.10.1 stringi_1.1.5   minqa_1.2.4
## [13] nloptr_1.0.4    Matrix_1.2-10   rmarkdown_1.6   splines_3.4.1
## [17] lme4_1.1-13     tools_3.4.1     stringr_1.2.0   yaml_2.1.14
## [21] compiler_3.4.1  htmltools_0.3.6 knitr_1.16
```