Cereal Investigation

By: Sharat Vyas, Chang Lu, Wei Zhang

12.04.2017

STAT 4123/5123

1.Abstract

Upon investigation we found many counts of highly correlated predictors. Once adjusting the model to take this into account, we found that shelf location and sugar have a relationship and as a result, we fit a interaction term to our model. Our final model found that cereal rating is significantly related with shelf, protein, fat, sodium, fiber, carbohydrates, sugars, and vitamins. We found that our model fit the data very well, it had an adjusted R^2 of 0.9932. The model also had an F-statistic of 890.6 with 12 and 61 degrees of freedom. Our residuals were normally distributed and centered around zero.

2.Introduction

Our group worked with the cereal.txt file from Carnegie Mellon University's Data and Story Library inside the Department of Statistics. The data came from a sample of 77 different breakfast cereals collected by workers. Using this data, we wanted to investigate what were the important factors that have the most profound effect on the overall rating of the cereal.

3.Description of Data

The dataset came with 15 dependent variables and 1 independent variables. The independent variable was response and the dependent variables consisted of 5 categorical predictors and 10 quantitative predictors. The 5 categorical predictors were cereal name, cereal manufacturer, cereal type(Hot/Cold), Vitamin percentage, and shelf location(1=Bottom, 2=Middle, 3=Top). The 10 quantitative predictors included the number of calories per serving, grams of protein, grams of fat, milligrams of sodium, grams of fiber, grams of carbohydrates, grams of sugar, milligrams of potassium, weight of one serving, and the number of cups in one serving. Upon exploratory analysis, we found multicollinearity between many of the predictors including fiber, potassium, calories and weight. The data also had 3 missing values, as a result we had to remove those observations from the dataset.

4.Analysis and Results

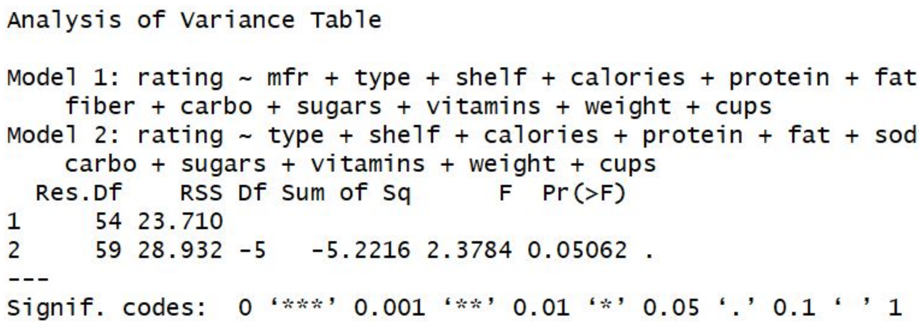
To begin, we performed exploratory data analysis. First, we dropped the name variable because it did not make sense to fit this on the response since name will not have any bearing on the overall health rating of the cereal. After this, we fit the full model and analyzed it. We found an error that the model had a perfect fit and this was an indicator for multicollinearity.

4.1 Addressing Multicollinearity

To address the multicollinearity, our group first constructed the hat matrix and analyzed the correlations between the predictors. We found that Potassium and Fiber were very highly correlated with a value of 0.912 which indicated almost perfect multicollinearity, as a result we decided to drop potassium from the model. We did not drop Fiber because we felt that consumers valued fiber content more than potassium content of cereal. After this, we also decided to drop the calorie predictor. We noticed that calories were highly correlated with weight. Upon further investigation, we found that fat was also correlated with calorie. This makes sense because fat content and weight both are factor in how many calories a specific food has.

4.2 Optimizing the Model

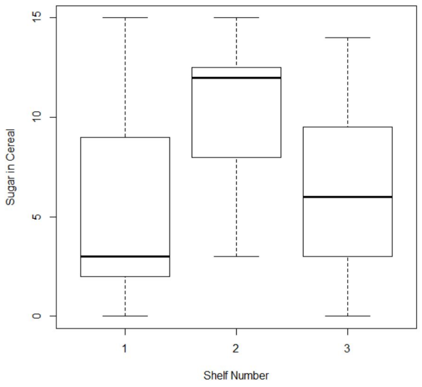
After we felt that multicollinearity was no longer an issue, we began to investigate what predictors were necessary. We fit the model that regressed manufacturer name, type, shelf, calories, protein, fat, sodium, fiber,carbohydrates ,sugars, vitamins, weight, and cups on rating. We looked at the model summary which indicated that manufacturer name may not be significant, as a result we fit another model without manufacturer name and compared the models to make sure that the coefficients did not drastically change. As a result, we dropped manufacturer name from the model. We performed similar tests for cups in a serving and cereal types, we found that they were both not significant and dropped them from the model. During our exploratory data analysis, we found there is likely some interaction between shelf location and sugar in cereal. We looked at boxplot and found that cereal on shelf 2 had a higher sugar content than cereals on shelves 1 or 3. This maybe because cereal on shelf 2 is likely directly in front of kids and as a result, it will appeal more to them. As a result we decided to add an interaction term to the model and see if this interaction is significant. We performed an anova test to compare 2 models, one with the interaction term and one model without. We found that the interaction term was significant, and as a result we decided to keep it in our model.



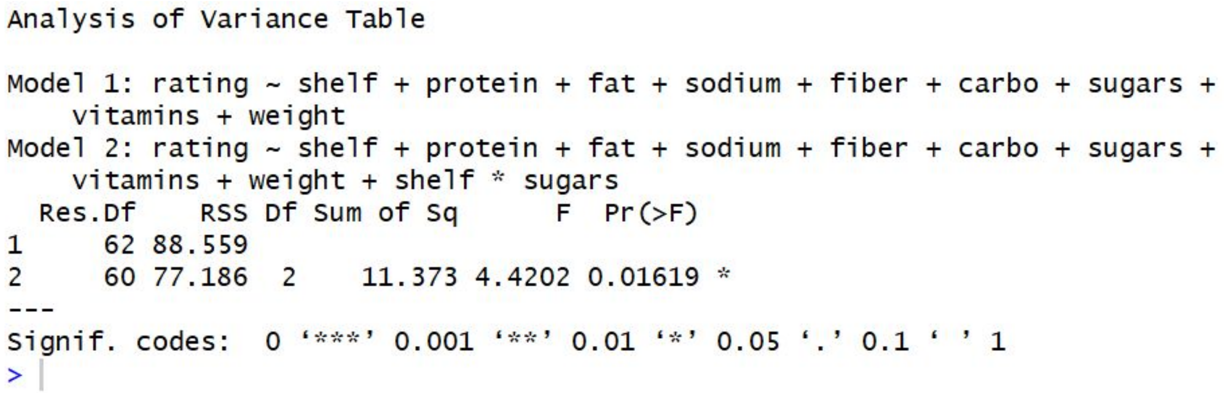
```
Analysis of Variance Table

Model 1: rating ~ mfr + type + shelf + calories + protein + fat + sodium +
    fiber + carbo + sugars + vitamins + weight + cups
Model 2: rating ~ type + shelf + calories + protein + fat + sodium + fiber +
    carbo + sugars + vitamins + weight + cups
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     54 23.710
2     59 28.932 -5   -5.2216 2.3784 0.05062 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Above, we see that manufacturer name is not significant at alpha=0.05 and as a result, drop it from our model

Figure A.1



From Figure A.1, we see the boxplot between shelf number and sugar content in cereal, it strongly indicates that shelf 2 has a higher sugar content.

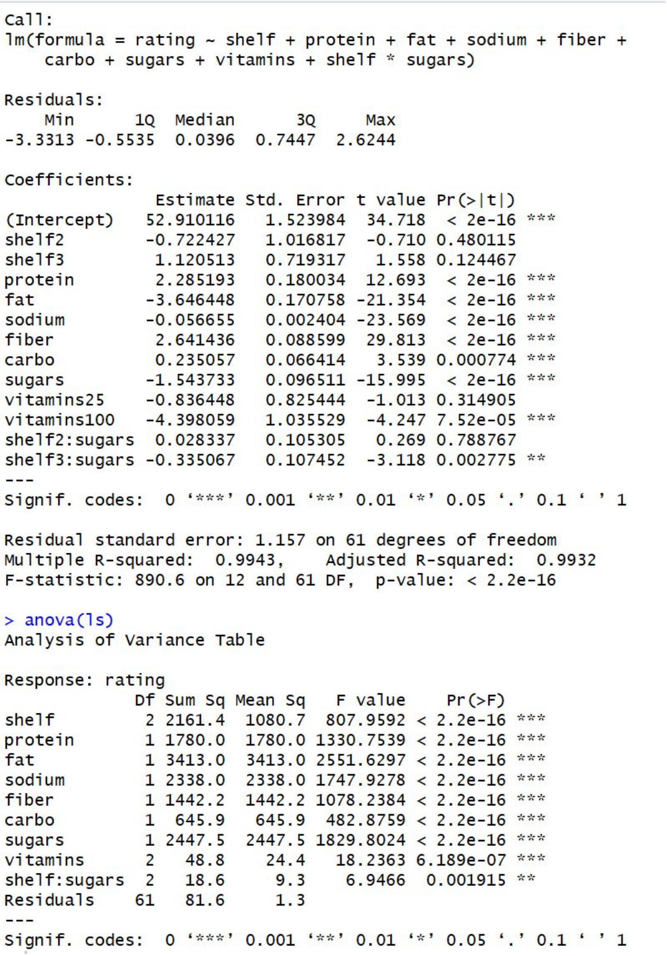


```
Analysis of Variance Table

Model 1: rating ~ shelf + protein + fat + sodium + fiber + carbo + sugars +
    vitamins + weight
Model 2: rating ~ shelf + protein + fat + sodium + fiber + carbo + sugars +
    vitamins + weight + shelf * sugars
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     62 88.559
2     60 77.186  2    11.373 4.4202 0.01619 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

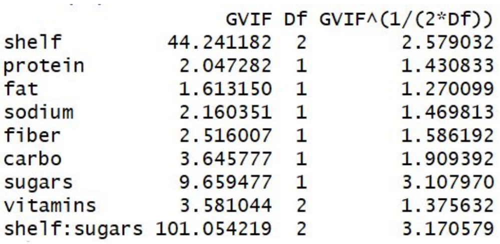
Above, we fit the interaction term between the shelf and sugar, we get a p-value of 0.01619 and using the same alpha=0.05, we find that the interaction term is in fact significant and decide to keep it in our model.

As a result, we fit the following model as our final model.

$$\widehat{Rating} = 52.91012 - 0.72243(shelf2) + 1.12051(shelf3) + 2.28519(protein) - 3.64645(fat) -0.05666(sodium) + 2.64144(fiber) + 0.23506(carbohydrates) - 1.54373(sugars) -0.83645(vitamins25) - 4.39806(vitamins100) + 0.02834(shelf2*sugars) -0.33507(Shelf3*sugars)$$



```
Call:
lm(formula = rating ~ shelf + protein + fat + sodium + fiber +
    carbo + sugars + vitamins + shelf * sugars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3313 -0.5535  0.0396  0.7447  2.6244

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   52.910116   1.523984  34.718  < 2e-16 ***
shelf2        -0.722427   1.016817  -0.710 0.480115
shelf3         1.120513   0.719317   1.558 0.124467
protein        2.285193   0.180034  12.693  < 2e-16 ***
fat           -3.646448   0.170758 -21.354  < 2e-16 ***
sodium        -0.056655   0.002404 -23.569  < 2e-16 ***
fiber          2.641436   0.088599  29.813  < 2e-16 ***
carbo          0.235057   0.066414   3.539 0.000774 ***
sugars        -1.543733   0.096511 -15.995  < 2e-16 ***
vitamins25    -0.836448   0.825444  -1.013 0.314905
vitamins100   -4.398059   1.035529  -4.247 7.52e-05 ***
shelf2:sugars  0.028337   0.105305   0.269 0.788767
shelf3:sugars -0.335067   0.107452  -3.118 0.002775 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.157 on 61 degrees of freedom
Multiple R-squared:  0.9943,	Adjusted R-squared:  0.9932
F-statistic: 890.6 on 12 and 61 DF,  p-value: < 2.2e-16

> anova(ls)
Analysis of Variance Table

Response: rating
             Df Sum Sq Mean Sq   F value    Pr(>F)
shelf         2 2161.4  1080.7  807.9592 < 2.2e-16 ***
protein       1 1780.0  1780.0 1330.7539 < 2.2e-16 ***
fat           1 3413.0  3413.0 2551.6297 < 2.2e-16 ***
sodium        1 2338.0  2338.0 1747.9278 < 2.2e-16 ***
fiber         1 1442.2  1442.2 1078.2384 < 2.2e-16 ***
carbo         1  645.9   645.9  482.8759 < 2.2e-16 ***
sugars        1 2447.5  2447.5 1829.8024 < 2.2e-16 ***
vitamins      2   48.8    24.4   18.2363 6.189e-07 ***
shelf:sugars  2   18.6     9.3    6.9466  0.001915 **
Residuals    61   81.6     1.3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

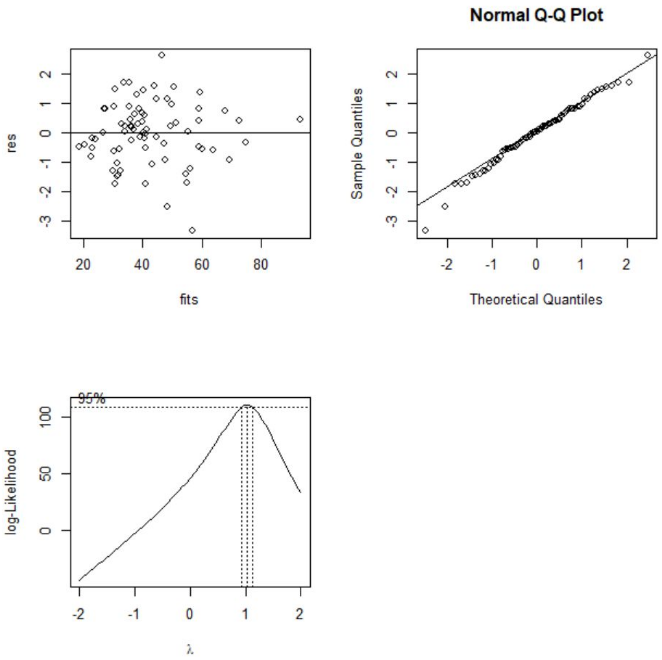


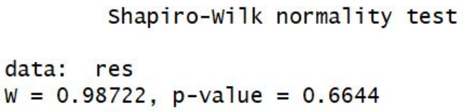
```
                   GVIF Df GVIF^(1/(2*Df))
shelf         44.241182  2        2.579032
protein        2.047282  1        1.430833
fat            1.613150  1        1.270099
sodium         2.160351  1        1.469813
fiber          2.516007  1        1.586192
carbo          3.645777  1        1.909392
sugars         9.659477  1        3.107970
vitamins       3.581044  2        1.375632
shelf:sugars 101.054219  2        3.170579
```

From the model summary above, we find that all the predictors are significant. We also see that we have a very high adjusted R^2 and that there is no longer an issue with multicollinearity. This has become our final model to be used in the residual analysis performed below.

4.3 Residual Analysis

Figure A.2





```
                Shapiro-Wilk normality test

data:  res
W = 0.98722, p-value = 0.6644
```

From the Normal Q-Q plot of figure A.2, we see that the residuals are normally distributed. The Shapiro-Wilk normality test, confirms our observation. The residuals have zero-mean and no pattern. The argument can be made that heteroskedasticity may occur because the residuals have a bowl shape. However, upon performing the Box-Cox transformation, we see that the suggested λ is 1 so no transformation is necessary. From the Q-Q plot, there is evidence of possible outliers and possible influential points. To investigate this, we standardized the residuals, calculated the leverages, and cook distances. We find 2 observations with a high leverage, but no points with a high cook distance which indicates that these point are not influential.

5.Discussion and Conclusions

From our working model, we can conclude that the rating is significantly related to shelf location, protein, fat, sodium, fiber, carbohydrates, sugars, vitamins. From our model we can conclude that cereals from shelf 2 tend to have a lower rating that cereals from shelf 1 and shelf 3. We found this to be very interesting. The difference between Shelf 2 and Shelf 1 was -0.72 +0.028(Sugar Content). The max sugar content in any cereal from the data was 15, as a result cereals from shelf 2 always has a lower rating than cereals from shelf 1 or 3. We also see that higher sugar, fat, sodium has a negative effect on a cereal's rating. This makes sense because a higher presence of these items makes a cereal less healthy. We found one interesting fact as well, cereals that have more vitamins tend to have a lower overall rating. Further investigation into why the higher vitamin content negatively impacts a cereal's overall rating is required. Having taste as one of the categorical predictors may help shed some light on this issue. It can improve the accuracy of the rating and can allow you to gather more information about how taste relates to vitamin content.

6.Technical Appendices

```
####Read in the Data##########

setwd("C:/Users/sidvy/Desktop/Stat4123Project")

data=read.table("cereal.txt", header=T)

data <- data[-c(5, 21, 58), ]

mfr=factor(data$mfr)

type=factor(data$type)

name=factor(data$name)

shelf=factor(data$shelf)

calories=data$calories

protein=data$protein

fat=data$fat

sodium=data$sodium

fiber=data$fiber

carbo=data$carbo

sugars=data$sugars
```

```
potass=data$potass

vitamins=factor(data$vitamins)

weight=data$weight

cups=data$cups

rating=data$rating


#######Exploratory Analysis#########

plot(shelf,sugars)



######Construct Correlation Matrix########

X=cbind(sugars,protein,fat,sodium,fiber,carbo,calories,weight,cups,potass,mfr,type,shelf,vitamins)

hat=X%*%solve(t(X)%*%X)%*%t(X)

diag(hat)

cor(X)

pairs(data)

#drop potass because of high correlation with fiber

ls2=lm(rating~mfr+shelf+calories+protein+fat+sodium+fiber+carbo+sugars+vitamins+weight+cups)

library(car)

vif(ls)


######fitting initial model############

ls=lm(rating~mfr+type+shelf+calories+protein+fat+sodium+fiber+carbo+sugars+vitamins+weight+cups)

ls2=lm(rating~type+shelf+calories+protein+fat+sodium+fiber+carbo+sugars+vitamins+weight+cups)

anova(ls,ls2)

#we decided to drop mfr using the anova test

ls=lm(rating~type+shelf+calories+protein+fat+sodium+fiber+carbo+sugars+vitamins+weight+cups)

ls2=lm(rating~type+shelf+protein+fat+sodium+fiber+carbo+sugars+vitamins+weight+cups)

anova(ls,ls2)

# we decide to drop calories because there is evidence of multicollinearity between weight…we see this because of a drastic change

# in the weight coefficient between the 2 models
```

```
ls=lm(rating~type+shelf+protein+fat+sodium+fiber+carbo+sugars+vitamins+weight+cups)

ls2=lm(rating~type+shelf+protein+fat+sodium+fiber+carbo+sugars+vitamins+weight)

anova(ls,ls2)

# we decide to drop cups from our anova test above


ls=lm(rating~type+shelf+protein+fat+sodium+fiber+carbo+sugars+vitamins+weight)

ls2=lm(rating~shelf+protein+fat+sodium+fiber+carbo+sugars+vitamins+weight)

anova(ls,ls2)

# we decide to drop type from our model using the anova test above


ls=lm(rating~shelf+protein+fat+sodium+fiber+carbo+sugars+vitamins+weight)

ls2=lm(rating~shelf+protein+fat+sodium+fiber+carbo+sugars+vitamins+weight+shelf*sugars)

anova(ls,ls2)

# we see that the interaction term between shelf and sugars is significant and decide to keep it


ls=lm(rating~shelf+protein+fat+sodium+fiber+carbo+sugars+vitamins+weight+shelf*sugars)

ls2=lm(rating~shelf+protein+fat+sodium+fiber+carbo+sugars+vitamins+shelf*sugars)

anova(ls,ls2)


# we decide to drop weight from the model using the anova test from above


ls=lm(rating~shelf+protein+fat+sodium+fiber+carbo+sugars+vitamins+shelf*sugars)

ls2=lm(rating~protein+fat+sodium+fiber+carbo+sugars+vitamins)

anova(ls,ls2)

#we see that shelf is signficant and decide to keep it as evident in the anova test above


####further analysis

par(mfrow=c(3,2))

plot(shelf,sugars)

plot(shelf,carbo)

plot(shelf,fat)

plot(shelf,protein)

plot(shelf,sodium)
plot(shelf,fiber)


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

ls=lm(rating~shelf+protein+fat+sodium+fiber+carbo+sugars+vitamins+shelf*sugars)

vif(ls)



#######Residual Analysis#######


###residual analysis of model1

res=ls$residuals

fits=ls$fitted.values

par(mfrow=c(2,2))

plot(fits, res)

abline(h=0)

qqnorm(res)

qqline(res)

shapiro.test(res)

result=boxcox(ls)

result$x[result$y==max(result$y)]

### outlier detection

#standardized residual

MSres=(summary(ls)$sigma)^2

stand=res/sqrt(MSres)

which(abs(stand)>3)

#leverage

#hat matrix

X=cbind(1,shelf,protein,fat,sodium,fiber,carbo,sugars,vitamins,shelf:sugars)

hat=X%*%solve(t(X)%*%X)%*%t(X)

lev=diag(hat)

lev

# cut-off point

p=dim(X)[2]
```

```
n=dim(X)[1]

2*p/n # leverage

which(lev>2*p/n)

#2,4,11,54,65

#cook's distance

cooks=cooks.distance(ls)

which(cooks>1) #none

# DEFIT

dffits=dffits(ls)

# cut-off point

2*sqrt(p/n) # for DFFITS

which(abs(dffits)>2*sqrt(p/n)) #4,7,11,29,43,44,54,65,68

# COVRATIO

# command

covratio=covratio(ls)

# cut-off point

1+3*p/n; 1-3*p/n # COVRATIO

which(covratio>1+3*p/n) #1,2,37,38,43,53,62,65,67,69

which(covratio<1-3*p/n) #8,32


####Model Done#######

ls=lm(rating~shelf+protein+fat+sodium+fiber+carbo+sugars+vitamins+shelf*sugars)

summary(ls)

anova(ls)

vif(ls)
```

7.Bibliography

“Healthy Breakfast.” *Healthy Breakfast Story*, lib.stat.cmu.edu/DASL/Stories/HealthyBreakfast.html.

“How Your Caloric Intake Affects Your Health.” *HealthStatus*, HealthStatus Team, 10 Jan. 2017, www.healthstatus.com/health_blog/body-fat-calculator-2/caloric-intake-affects-health-2/.