The Multiple Linear Regression of Cereal Ratings

Cereal Ratings have long been a rating system by the Healthy Cereals Organization (HCO). The HCO rates breakfast cereals from 0-100, but how it comes up with this score has remained a mystery. Thus, the path to uncovering this mystery was to look at nutritional facts of cereals in the following categories as prediction variables: protein, sodium, fiber, sugars, vitamins, weight, and cups. A multiple linear regression was performed with cereal rating as the response variable.

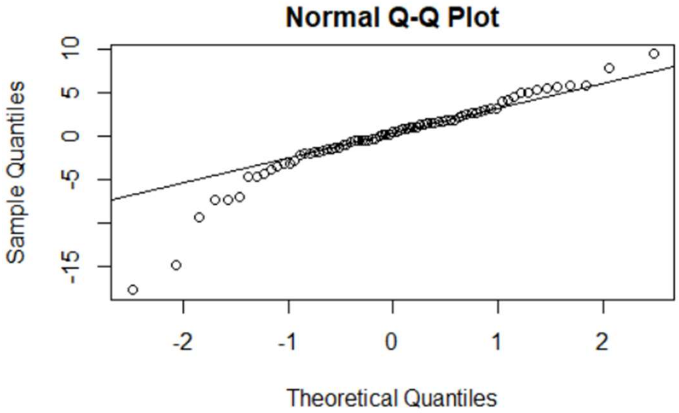
$$Rating_i = \beta_0 + \beta_1 protein_i + \beta_2 sodium_i + \beta_3 fiber_i + \beta_4 sugars_i + \beta_5 vitamins_i + \beta_6 weight_i + \beta_7 cups_i$$

The null hypothesis was that neither of these variables had any influence on cereal rating, however.

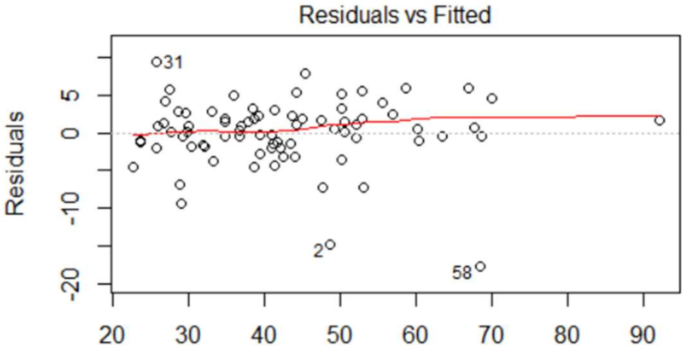
$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$$

$$H_1: at\ least\ one\ \beta_j \neq 0\ with\ j\ in\ the\ range = [1,7]$$

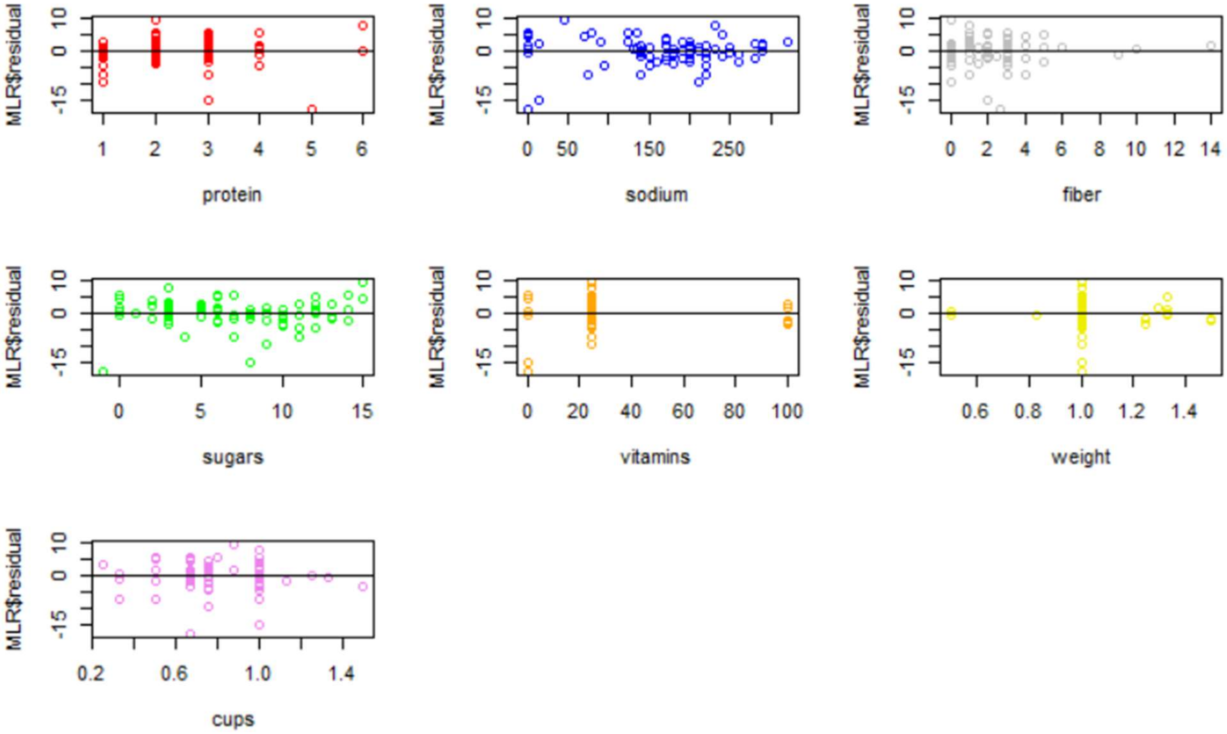
The model passed the four checks of a linear regression. The Q-Q Plot below shows a normality check. The points follow the trend line very closely except for the bottom of the left tail.



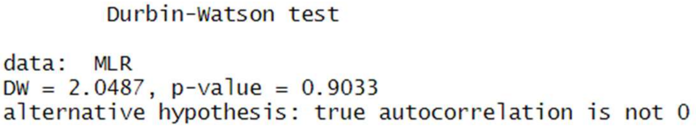
The Residuals vs Fitted Plot shows is a check on equal variance. There does not appear to be a trend at all with the data points.



The model also passes a linearity check with Residuals vs. Fitted plots of protein, sodium, fiber, sugars, vitamins, weight, and cups.

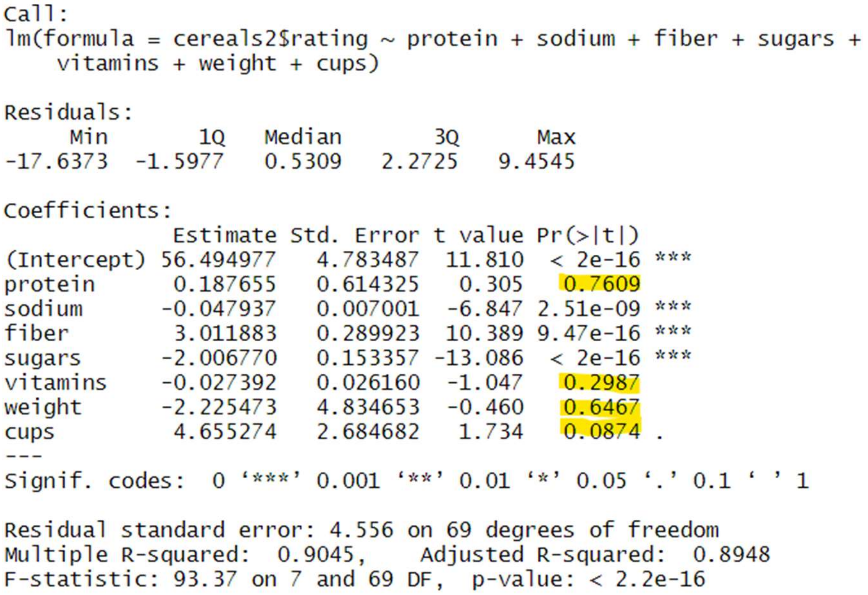


Finally, the Durbin-Watson test secured the ratings and errors were independent. The p-value of 0.9033 shows this.


```
        Durbin-Watson test

data:  MLR
DW = 2.0487, p-value = 0.9033
alternative hypothesis: true autocorrelation is not 0
```

However, the model could be simplified since when running a t-test on each prediction variable, the p-values for some were not below the selected alpha of 0.05. They are highlighted.

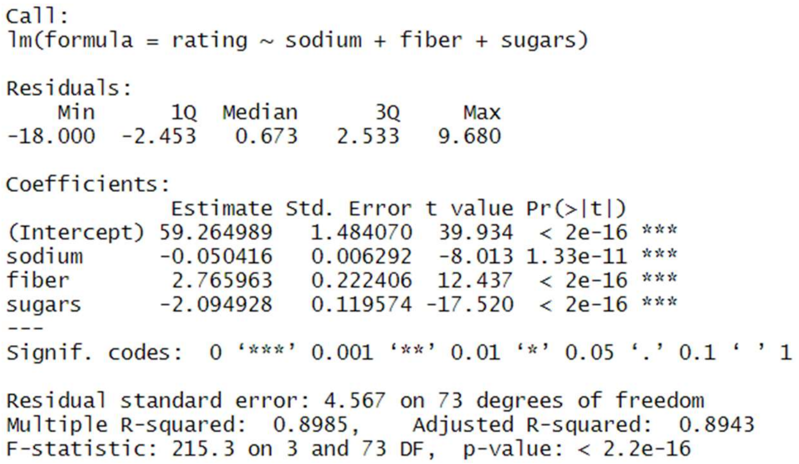

```
Call:
lm(formula = cereals2$rating ~ protein + sodium + fiber + sugars + 
    vitamins + weight + cups)

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 ***
protein      0.187655   0.614325   0.305   0.7609    
sodium      -0.047937   0.007001  -6.847 2.51e-09 ***
fiber        3.011883   0.289923  10.389 9.47e-16 ***
sugars      -2.006770   0.153357 -13.086  < 2e-16 ***
vitamins    -0.027392   0.026160  -1.047   0.2987    
weight      -2.225473   4.834653  -0.460   0.6467    
cups         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
```

Since, the F-statistic was 93.37, the null hypothesis was rejected since in the p-value was much less than alpha which equaled 0.05. Protein, vitamins, weight, and cups were one-by-one removed from the model using backwards selection/elimination. When this was done, the final results still kept the four checks on linearity, independence of ratings and errors, normality of the model, and equality of variances. As you can see, the adjusted R-squared remained the same in the new model, meaning protein, vitamins, weight, and cups did not have much influence on the model at all.



```
Call:
lm(formula = rating ~ sodium + fiber + sugars)

Residuals:
    Min      1Q  Median      3Q     Max 
-18.000  -2.453   0.673   2.533   9.680 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 59.264989   1.484070  39.934  < 2e-16 ***
sodium      -0.050416   0.006292  -8.013 1.33e-11 ***
fiber        2.765963   0.222406  12.437  < 2e-16 ***
sugars      -2.094928   0.119574 -17.520  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.567 on 73 degrees of freedom
Multiple R-squared:  0.8985,	Adjusted R-squared:  0.8943 
F-statistic: 215.3 on 3 and 73 DF,  p-value: < 2.2e-16
```

Thus, the best model for predicting Cereal Ratings from the HCO is:

$$Rating_i \; = \; \beta_0 \; - \; \beta_1 sodium_i + \beta_2 fiber_i \; - \; \beta_3 sugar_i$$

The Cereal Ratings can now be exploited into the breakfast cereal industry’s favor by just lowering sodium, increasing fiber, and lowering sugar.

R-Code:

```
cereals = read.csv("CerealsRating.csv")
str(cereals)
summary(cereals)
cereals2 <- cereals[,2:length(cereals)]

## Normality Check
attach(cereals2)
MLR <- lm(rating~protein+sodium+fiber+sugars+vitamins+weight+cups)
summary(MLR)
qqnorm(MLR$residual)
qqline(MLR$residual)
shapiro.test(MLR$residual)

## Equal Variance Check
plot(MLR)

#The points on the residuals vs fitted are skewed to the right
```

```
## Linearity Check
par(mfrow=c(3,3))
plot(x = protein, y = MLR$residual, col = "red")
abline(h=0)
plot(x = sodium, y = MLR$residual, col = "blue")
abline(h=0)
plot(x = fiber, y = MLR$residual, col = "gray")
abline(h=0)
plot(x = sugars, y = MLR$residual, col = "green")
abline(h=0)
plot(x = vitamins, y = MLR$residual, col = "orange")
abline(h=0)
plot(x = weight, y = MLR$residual, col = "yellow2")
abline(h=0)
plot(x = cups, y = MLR$residual, col = "violet")
abline(h=0)

## Independence Check (Durbin-Watson): Not required since not time-series data
dwtest(MLR, alternative = "two.sided")

## Transformation on Y
library(MASS)
trans <- boxcox(rating~protein+sodium+fiber+sugars+vitamins+weight+cups,
          data=cereals2,
          plotit=F, #get boxcox instead of the plot
          lambda = seq(-3, 3, by=0.125))
maxyentry <- which.max(trans$y) # obtain index number of largest y
trans$x[maxyentry] # obtain lambda value

## Add new column to dataframe with transformed variable
cereals2$logY <- log(cereals2$rating)
cereals2$logprotein <- log(cereals2$protein)
cereals2$logsodium <- log(cereals2$sodium)
cereals2$logfiber <- log(cereals2$fiber)
cereals2$logvitamins <- log(cereals2$vitamins)
cereals2$logweight <- log(cereals2$weight)

## Full Model with Y-Transformation
library(car)
attach(cereals2)
MLR2 <- lm(logY~protein+sodium+fiber+sugars+vitamins+weight+cups)
summary(MLR2) # gives partial t-test values and regression F value
anova(MLR2) # provides ANOVA table for overall model

## Normality Check
par(mfrow=c(1,1))
qqnorm(MLR2$residual)
qqline(MLR2$residual)
```

```
shapiro.test(MLR2$residual)

## Equal Variance Check (top left corner)
par(mfrow=c(2,2))
plot(MLR2)

## Linearity Check
par(mfrow=c(3,3))
plot(x = protein, y = MLR2$residual, col = "red")
abline(h=0)
plot(x = sodium, y = MLR2$residual, col = "blue")
abline(h=0)
plot(x = fiber, y = MLR2$residual, col = "gray")
abline(h=0)
plot(x = sugars, y = MLR2$residual, col = "green")
abline(h=0)
plot(x = vitamins, y = MLR2$residual, col = "orange")
abline(h=0)
plot(x = weight, y = MLR2$residual, col = "yellow2")
abline(h=0)
plot(x = cups, y = MLR2$residual, col = "violet")
abline(h=0)

## VIF Check
vif(MLR2)

MLR3 <- lm(rating~protein+sodium+fiber+sugars+vitamins+cups)
summary(MLR3)
MLR4 <- lm(rating~protein+sodium+fiber+sugars+cups)
summary(MLR4)
MLR5 <- lm(rating~protein+sodium+fiber+sugars)
summary(MLR5)
anova(MLR5)
MLR6 <- lm(rating~sodium+fiber+sugars)
summary(MLR6)
anova(MLR6)

#Normality Check
qqnorm(MLR6$residual)
qqline(MLR6$residual)
shapiro.test(MLR6$residual)

## Equal Variance Check
plot(MLR6)

## Linearity Check
        par(mfrow=c(1,3))
```

```
plot(x = cereals2$sodium, y = MLR6$residual, col = "blue")
abline(h=0)
plot(x = cereals2$logfiber, y = MLR6$residual, col = "gray")
abline(h=0)
plot(x = sugars, y = MLR6$residual, col = "green")
abline(h=0)

## Independence Check (Durbin-Watson): Not required since not time-series data
dwtest(MLR, alternative = "two.sided")
```