An MLR and ANOVA Approach to Predicting Cereal Ratings

Ilan Shlain

Analysis Summary

In this project, we investigated how certain nutritional characteristics of cereal contributed to their overall given consumer rating. Using a data set with 77 entries, we fitted a multiple linear regression model with rating as the response variable, and 7 predictors. The 7 predictors were: protein, sodium, fiber, sugars, vitamins, weight, and cups per serving. An Exploratory Data Analysis was conducted to understand the distribution of data, identify potential outliers, and observe any correlation between variables. A full multiple regression model was fit and then evaluated using ANOVA, diagnostic plots, and variance inflation factors to test model assumptions and multicollinearity. Through backward variable selection, a final parsimonious model was selected that retained sodium, fiber, and sugars as statistically significant predictors. The results indicate that higher fiber content is associated with higher ratings, while higher sugar and sodium levels are associated with lower ratings, with sugar exhibiting the strongest negative association. Diagnostic checks confirmed that assumptions of linearity and constant variance were reasonably satisfied; while mild non-normality was observed in the residual tails, multicollinearity was not a major concern. Ultimately, the analysis suggests that consumer ratings favor cereals that are high in fiber and low in sugar and sodium.

Data Introduction and Despcription

The original dataset contains 77 observations and 16 variables. For the purposes of this analysis, we focus on the response variable (rating) and seven relevant predictors. Observations containing negative values, which indicate missing data in this dataset, were removed prior to analysis.

Table 1: First 5 rows of our data set.

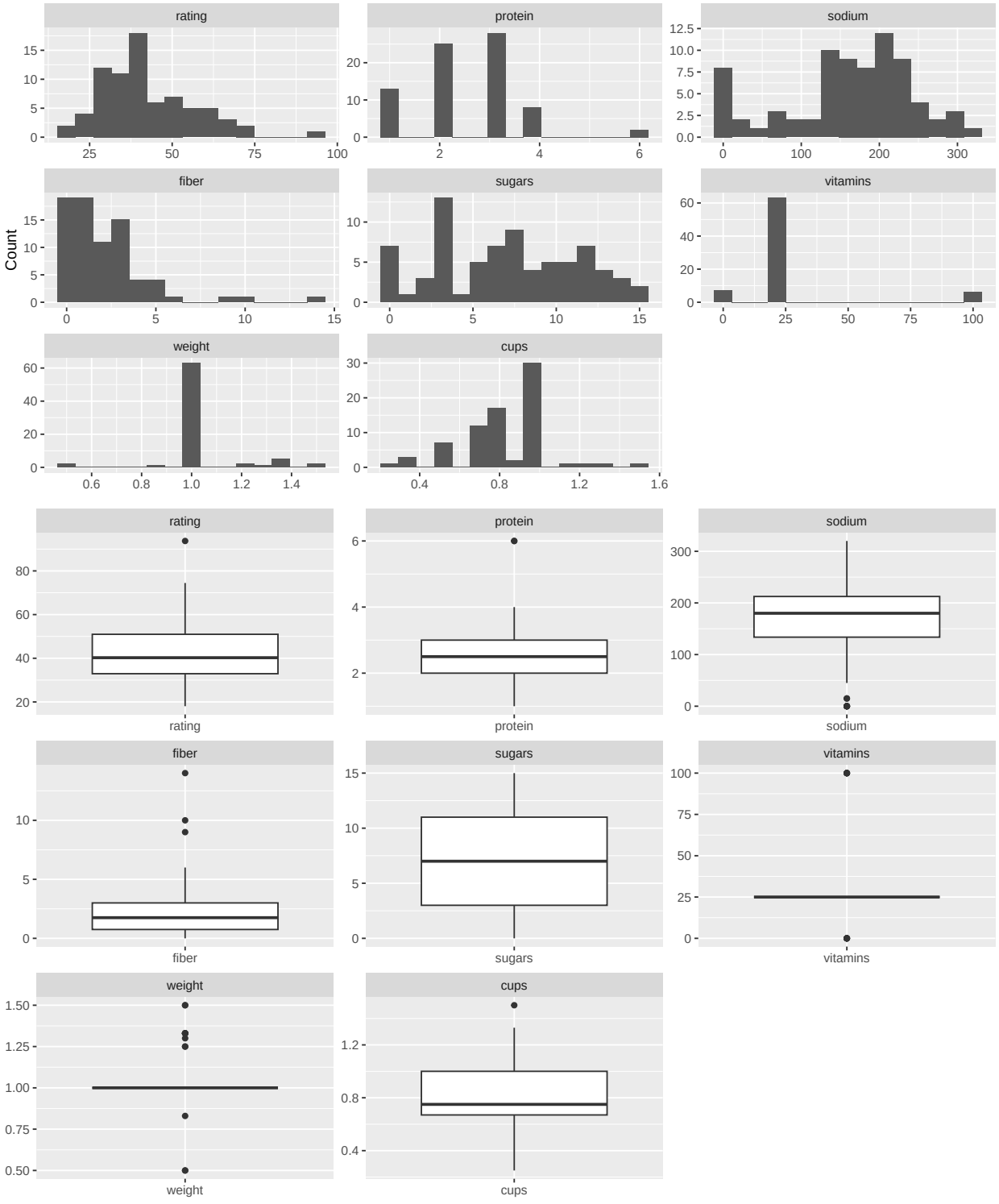
rating	protein	sodium	fiber	sugars	vitamins	weight	cups
68.403	4	130	10	6	25	1	0.33
33.984	3	15	2	8	0	1	1.00
59.426	4	260	9	5	25	1	0.33
93.705	4	140	14	0	25	1	0.50
34.385	2	200	1	8	25	1	0.75

Below, you can see a descriptive statistics summary table. From the table we can see that the variables exhibit differing scales and variability, with wide ranges observed in both the response and several nutritional predictors, motivating further regression analysis.

```
##                Min        Q1    Median        Mean        Q3       Max         SD
## rating    18.04285  32.93247  40.25309  42.5583012  50.9718  93.70491 14.1087652
## protein    1.00000   2.00000   2.50000   2.5131579   3.0000   6.00000  1.0644989
## sodium     0.00000 133.75000 180.00000 161.7763158 212.5000 320.00000 82.3233622
## fiber      0.00000   0.75000   1.75000   2.1447368   3.0000  14.00000  2.3983547
## sugars     0.00000   3.00000   7.00000   7.0263158  11.0000  15.00000  4.3786564
## vitamins   0.00000  25.00000  25.00000  28.6184211  25.0000 100.00000 22.2500739
## weight     0.50000   1.00000   1.00000   1.0300000   1.0000   1.50000  0.1514376
## cups       0.25000   0.67000   0.75000   0.8230263   1.0000   1.50000  0.2336038
```

Box Plots and Histograms

Here we display histograms and boxplots for the response (rating) and the seven predictors. Several variables show skewness and potential outliers (e.g., fiber, sodium, vitamins), which motivates checking regression assumptions and influential points in later sections.



Multiple Linear Regression and ANOVA Results

We model cereal rating as a linear function of the selected nutritional predictors using multiple linear regression (MLR). This approach allows us to assess the marginal effect of each predictor on the response while controlling for the others.

$$\textbf{Model:} \quad Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n.$$

$$\textbf{Assumptions:} \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2),$$
$$E(\varepsilon_i) = 0, \qquad \text{Var}(\varepsilon_i) = \sigma^2, \qquad \text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \ (i \neq j).$$

Test Hypotheses:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \qquad \text{vs} \qquad H_A : \text{at least one } \beta_j \neq 0$$

ANOVA

Table 2: ANOVA table for the full MLR model.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
protein	1	3302.1658	3302.1658	212.1409	0.0000
sodium	1	2367.9759	2367.9759	152.1258	0.0000
fiber	1	2037.8764	2037.8764	130.9192	0.0000
sugars	1	6080.7729	6080.7729	390.6468	0.0000
vitamins	1	35.6342	35.6342	2.2892	0.1349
weight	1	0.3414	0.3414	0.0219	0.8827
cups	1	46.0457	46.0457	2.9581	0.0900
Residuals	68	1058.4818	15.5659	NA	NA

Rejection of the Overall Null Hypothesis

The ANOVA results indicate that the overall model is statistically significant, providing evidence that at least 1 nutritional predictor has an association with cereal rating. Protien, sodium, fiber, and sugars show strong statistical signifigance, while vitamins, weigth, and cups were not deemed significant by the model at the 10% signifigance level. These results motivate further variable selection for model improvement.

Checking Model Assumptions

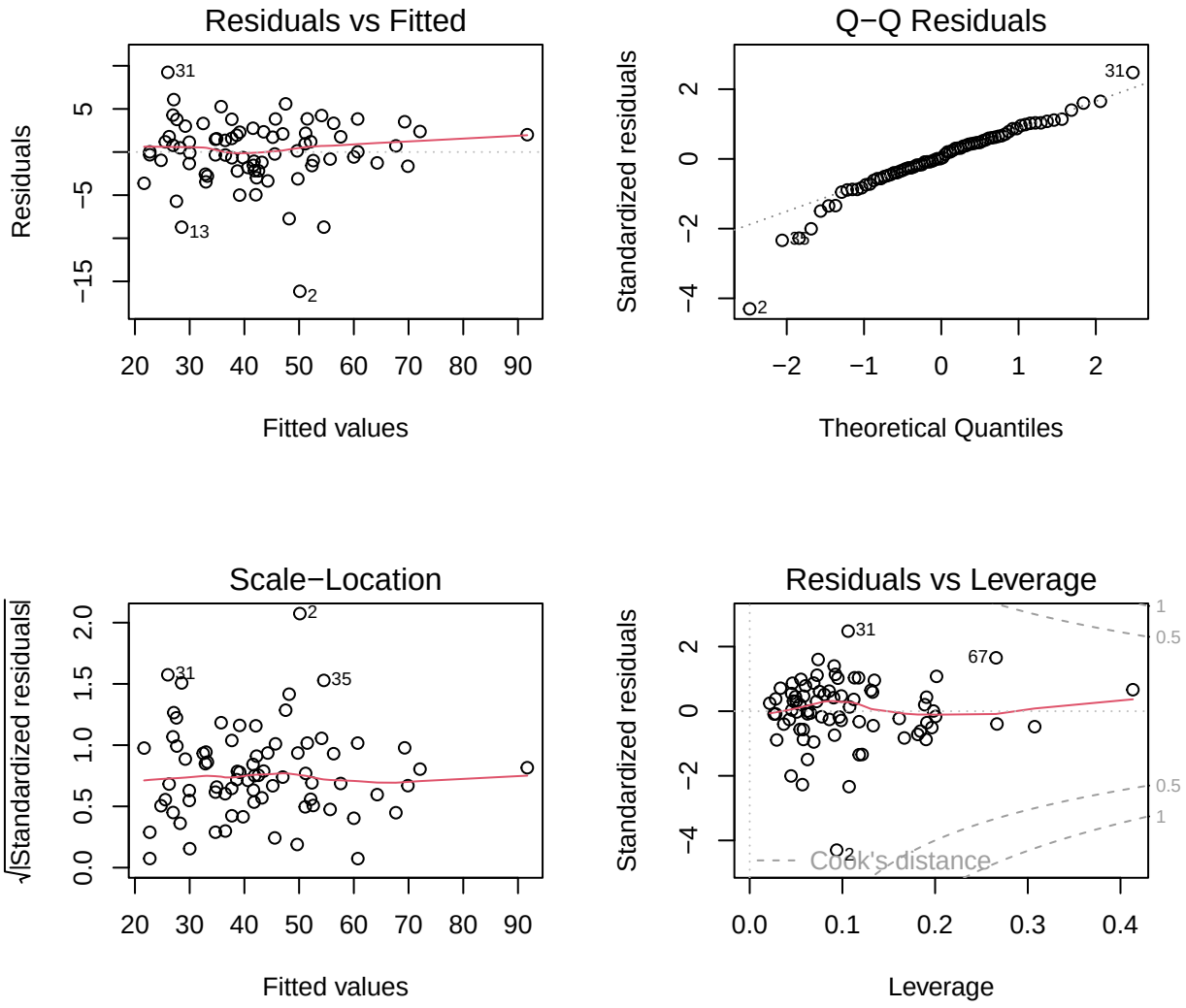


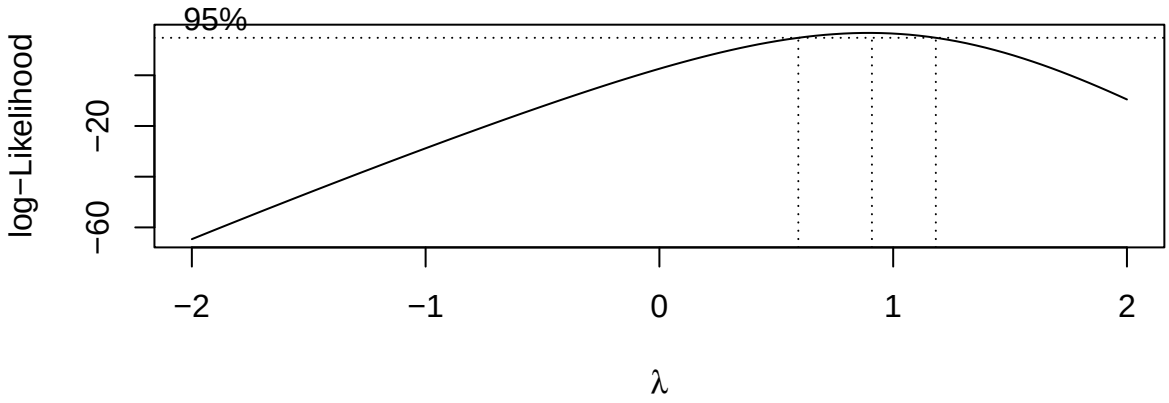
Table 3: Shapiro-Wilk normality test (residuals).

W	p_value
0.9277	3e-04

The residuals versus fitted plot shows no strong pattern, supporting linearity. The scale-location plot suggest approxinmatley constant variance. However the Q-Q plot and Shapiro-Wilk test shows a departure from normality. Residuals vs leverage identify several observations with higher influence, motivating further assesment of their impact.

Additional Diagnostics: Box–Cox and VIF Analysis

Normality diagnostics indicate a departure from normality in the residual tails. Although the Shapiro–Wilk test is statistically significant, the Q–Q plot suggests that deviations are primarily confined to the tails, with the center of the distribution remaining approximately linear. Since the Box–Cox analysis yielded an estimated λ close to 1 and regression is robust to moderate non-normality, the model was retained on the original scale.



Multicolinearity Analysis VIF

Table 4: Variance Inflation Factors (VIF) for the full MLR model.

	Predictor	VIF
6	weight	1.960
3	fiber	1.798
4	sugars	1.669
1	protein	1.611
7	cups	1.426
2	sodium	1.247
5	vitamins	1.235

The variance inflation facotors (VIF's) for all predictors are below 2, which is well under the usual threshold of 5 for problematic multicollinearity. This indicates that the predictors do not show strong linear dependance and that the estimated regression coefficients are not inflated due to colinearity. Therefore colinearity does not pose a major concern for the MLR model. Also, Cook's distance diagnostics indicate a small number of moderately influential observations, though none exceed standard thresholds for undue influence. As a result, all observations were retained for further analysis.

Backwards Variable Selection

In this section, we build the "best" multiple linear regression model using backward selection. We begin with the full model containing all seven predictors and iteratively remove the least significant predictor until all remaining predictors have p-values 0.10. This approach favors a more parsimonious model while maintaining explanatory power.

Table 5: Backward elimination steps (remove if p > 0.10).

step	removed	p_value
1	weight	0.9943
2	protein	0.1619
3	vitamins	0.1179
4	cups	0.1322

Table 6: Final model coefficient summary.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	61.0841	1.3631	44.8128	0
sodium	-0.0559	0.0057	-9.8683	0
fiber	2.7525	0.1959	14.0506	0
sugars	-2.1904	0.1072	-20.4233	0

Table 7: ANOVA table for final selected model.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sodium	1	2354.602	2354.6017	145.5208	0
fiber	1	4660.651	4660.6515	288.0409	0
sugars	1	6749.044	6749.0436	417.1092	0
Residuals	72	1164.997	16.1805	NA	NA

The final selected model retains sodium, fiber, and sugars as statistically significant predictors of cereal rating. Holding other variables constant, higher fiber content is associated with higher ratings, while higher sugar and sodium contents are associated with lower ratings. These effects are both statistically significant and practically meaningful, indicating that nutritional composition plays a key role in determining cereal rating.

Final Model Summary and Conclusion

The final selected multiple linear regression model includes sodium, fiber, and sugars as predictors of cereal rating. This model was obtained through backward elimination and represents a parsimonious specification that retains only statistically significant nutritional variables.

$$\widehat{\text{rating}} = 61.08 - 0.056(\text{sodium}) + 2.75(\text{fiber}) - 2.19(\text{sugars})$$

Real World Interpreation

The model results indicate that cereals that are higher in fiber content tend to be given higher ratings, while cereals with higher sugar and sodium content are given lower ratings. From the retained predictors, sugar exhibits the strongest neggative association with rating, followed by sodium. These results seem to align with nutritional expectation and suggest that consumer ratings favor cerals high in fiber and low in sugar and sodium.

Model Validity Recap

Model diagnostics suggest that the assumptions of linearity and constant variance are reasonably satisfied. While mild departures from normality were observed in the residual tails, these were not severe and were not improved by transformation, as seen by the Box–Cox analysis. Multcollinearity and influential observations were assessed and found not to pose major concerns, supporting the validity of the final model.

Limitations

This analysis is limited by sample size and the exclusion of non-nutritional factors that may influence cereal ratings. Although residuals show mild non-normality in the tails, regression diagnostics indicate no major violations affecting inference. Conclusions made from this model should take into account this constraints.

Appendix

The following code is in R and is what was used for this analysis. Displayed tables look slightly different in aesthetic, but that is only due to compilation and the use of knitr() and kable() functions.

```r
library(readr)
library(dplyr)
library(tidyr)
library(ggplot2)
library(car)
library(MASS)

CerealsRating <- read_csv("CerealsRating.csv")

predictors <- c("protein", "sodium", "fiber", "sugars", "vitamins", "weight", "cups")

df <- CerealsRating %>%
  dplyr::select(all_of(c("rating", predictors))) %>%
  mutate(across(everything(), ~ na_if(.x, -1))) %>%
  mutate(across(everything(), as.numeric)) %>%
  drop_na()

df <- as.data.frame(df)

# Five-number summary
summary_tbl <- data.frame(
  Min    = apply(df, 2, min),
  Q1     = apply(df, 2, quantile, 0.25),
  Median = apply(df, 2, median),
  Mean   = apply(df, 2, mean),
  Q3     = apply(df, 2, quantile, 0.75),
  Max    = apply(df, 2, max),
  SD     = apply(df, 2, sd)
)

# Histograms and boxplots
df_long <- stack(df)

ggplot(df_long, aes(x = values)) +
  geom_histogram(bins = 15) +
  facet_wrap(~ind, scales = "free")

ggplot(df_long, aes(x = ind, y = values)) +
  geom_boxplot() +
  facet_wrap(~ind, scales = "free")

# Full MLR model
```

```r
model_full <- lm(rating ~ protein + sodium + fiber + sugars + vitamins + weight + cups, data =
summary(model_full)
anova(model_full)

# Diagnostics
par(mfrow = c(2,2))
plot(model_full)
par(mfrow = c(1,1))

residualPlots(model_full)
shapiro.test(residuals(model_full))

# Box-Cox
boxcox(model_full)

# VIF
vif(model_full)

# Outliers and influence
r <- rstudent(model_full)
which(abs(r) > 3)

cooks <- cooks.distance(model_full)
which(cooks > 4/length(cooks))

# Backward selection
current <- model_full
repeat {
  pvals <- summary(current)$coefficients[-1, 4]
  if (max(pvals) <= 0.10) break
  worst_var <- names(which.max(pvals))
  current <- update(current, paste(". ~ . -", worst_var))
}

model_final <- current
summary(model_final)
anova(model_final)

par(mfrow = c(2,2))
plot(model_final)
par(mfrow = c(1,1))
shapiro.test(residuals(model_final))
```