Multiple Linear Regression:
What Makes a Good Cup of Coffee

Jeanette Nguyen
STAT 632: Linear and Logistic Regression
Dr. Joshua Kerr
May 9th, 2023

Introduction

During the pandemic, I got into ‘specialty coffee’. ‘Specialty coffee’ is used to refer to coffee that is graded 80 points or above on a 100 point scale by a certified coffee taster. It led me to wonder ‘what makes a good cup of coffee’ and ‘how is it graded’?

In order to find out what makes a good cup of coffee, the Multiple Linear Regression (MLR) model will be used. The goals of this study are 1) to estimate the relationship between the response variable, Total Cupping Points (the result of a graded coffee cup), and the explanatory variables (Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity, Cup Cleanliness, Sweetness, and Cupper Points) and 2) to determine the most/least important factors for a given value of the Total Cupping Points. The null hypothesis is that there is no useful linear relationship between Total Cupping Points and any of our predictors (Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity, Cup Cleanliness, Sweetness, and Cupper Points). The alternative hypothesis is that there is at least one useful relationship between the response variable and predictor variables.

Data Description

The dataset used for this study is collected from the Coffee Quality Institute's review pages in January 2018 of reviews of 1340 Arabica and Robusta coffee beans. Before applying the data, coffee beans that were missing Total Cupping scores were necessarily removed, resulting in our study using a total of 1338 coffee bean scores instead. We can assume the data was collected using a voluntary sampling method as any of the Coffee Quality Institute's trained reviewers are able to give a review and score to the coffee beans that they receive samples of. To determine the coffee quality, we will look at total cupping points(response variable) which are scored from 1 to 100. Total cupping points factor in variables such as Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity, Cup Cleanliness, Sweetness, Cupper Points, respectively, on a scale of 1 to 10 rating to contribute to the total cupping score.

We will use Total Cupping Points (response variable) and ten predictor variables (Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity, Cup Cleanliness, Sweetness, Cupper Points) to conduct a Multiple Linear Regression analysis.

Below are some summary statistics from the data which shows us that the average Total Cupping Score is 82.1512 with a standard deviation of 2.669 and the average mean for our predictor variables Aroma are around 7.5, with the exception of Uniformity, Cup Cleanliness, and Sweetness to average around 9.8. At first glance, I would assume that Uniformity, Cup

Cleanliness, and Sweetness to be the most important variables in determining Total Cupping Points.

Variables	**Mean**	**SD**
Total Cupping Points	82.1512	2.669
Aroma	7.5724	0.3159
Flavor	7.5261	0.3414
Aftertaste	7.4066	0.3503
Acidity	7.5413	0.3192
Body	7.5231	0.3078
Balance	7.5236	0.3536
Uniformity	9.8422	0.4852
Cup Cleanliness	9.8423	0.7153
Sweetness	9.8640	0.5542
Cupper Points	7.508	0.4268

Table 1: Summary Statistics

In the following page is a scatterplot matrix which shows the relationships between Total Cupping Points and each of the independent predictor variables respectively: Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity, Cup Cleanliness, Sweetness, Cupper Points. For the most part, it looks like there is a positive correlation between the response variable and the predictor variables.

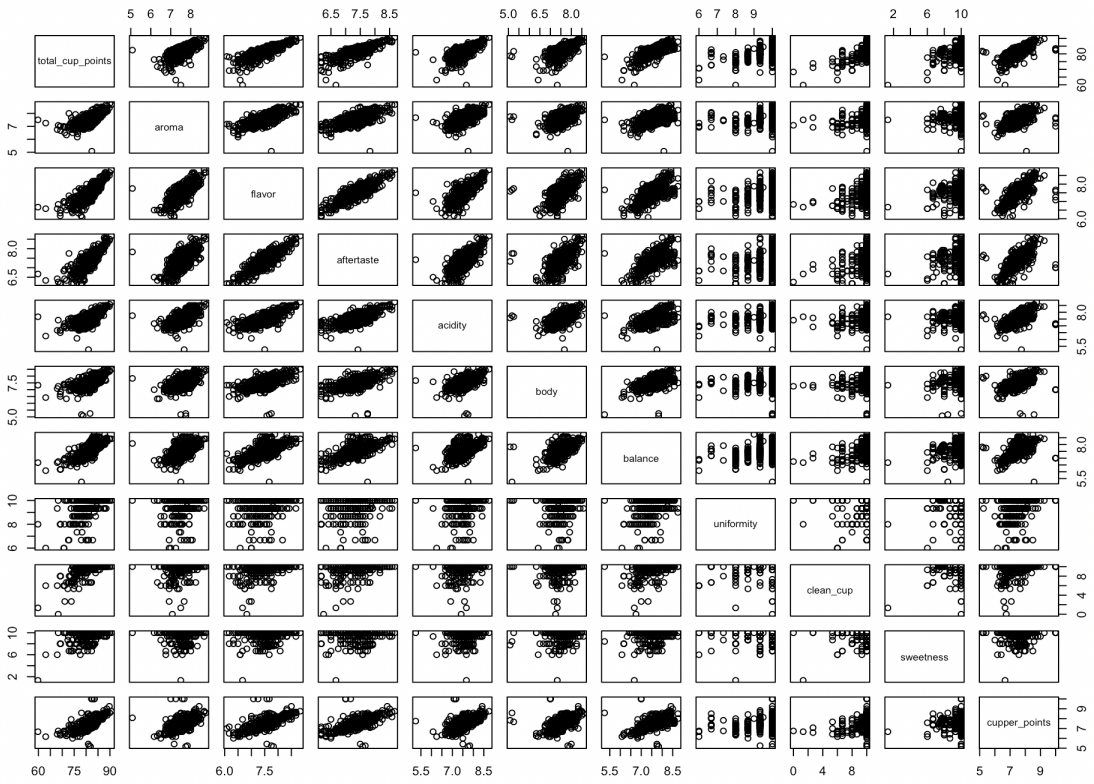


Figure 1: Scatter plot matrix

Methods and Results

Our model will take the form of $Y = \beta0 + \beta1X1 + \beta2X2 + \beta3X3 + ... + \beta nXn + \varepsilon i$, where $\beta0$ is our constant term and the y-intercept which will tell us the estimated average Total Cupping Points for coffee beans with our predictor units = 0. $\beta1$ to $\beta10$ represents the slope which will tell us the average difference in total cupping points for coffee beans whose predictor variable differs by one unit. X1 to X10 being our predictor variables respectively. εi accounts for random error. Y is our response variable, Total Cupping Points.

Using the same scatterplot matrix above, the conditions for linearity for the MLR model can be satisfied. Other conditions such as equal variance and normality are satisfied as most of the points are on the line for the Residuals vs Fitted plot (see appendix A) and also in the QQ plot (see appendix B). When checking for multicollinearity with a correlation matrix (see appendix C), the variable Flavor had a high correlation coefficient of 0.860, which suggests that the variable should be removed from the model for better interpretability. This tests the null hypothesis (H0), where the variables that we removed previously have no significance, against the alternative hypothesis (H1) that those variables are significant. However, running a reduced model against the full model resulted in a p-value of $< 2.2e\text{-}16$ (see appendix D), which we can reject the null hypothesis and conclude that Flavor is significant and we keep it in our model.

With our conditions tested, our estimated model is:
^Total Cupping Points = -0.0275 + 1.0006*Aroma + 0.9989*Flavor + 1.0024*Aftertaste + 0.9982*Acidity + 1.0012*Body + 1.0025*Balance + 1.0028*Uniformity + 1.0005*Cup Cleanliness + 0.9991*Sweetness + 0.9968*Cupper Points

Conclusion

From our modeling results (see Appendix E), the p-value is < 2.2e-16 is much smaller than our .05 level of significance so therefore, the null hypothesis is rejected. It is concluded that all coefficient β's are significant and have a correlation with the total cupping score. From our findings we find that our adjusted R^2 is 1, which implies that the model is the perfect fit. The most reliable predictor for predicting Total Cupping Score turned out to be Cup Cleanliness and the least reliable predictor was Flavor.

Some of the limitations in this study are that our results were an adjusted R^2 of 1. The fact that our variables fit perfectly does not account for the defects of a coffee bean that detract from a score. There is also the fact that Flavor seems to have correlation with aftertaste, which makes it multicollinear. Another concern is that Cupper Points is based on the cupper's (person who tastes) own preference and bias in the rating. Similarly, although reviewers are trained, certified, and giving a blind evaluation of the coffee sample received, there is only one reviewer for that coffee bean sample, and may contain bias even if it is a systematic process.

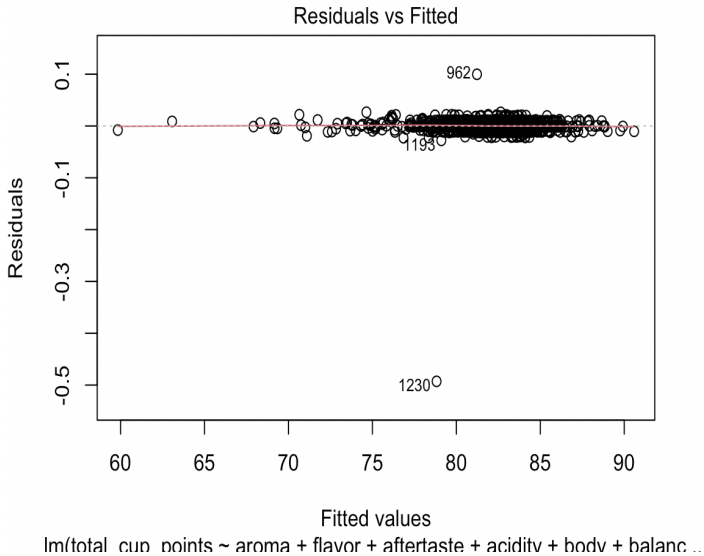
References

LeDoux, James, Coffee ratings (2020), GitHub repository, https://github.com/rfordatascience/tidytuesday/tree/master/data/2020/2020-07-07#coffee_ratings csv

"What Is Specialty Coffee." *The Specialty Coffee Company,* www.thespecialtycoffeecompany.com/resources/specialty-coffee/

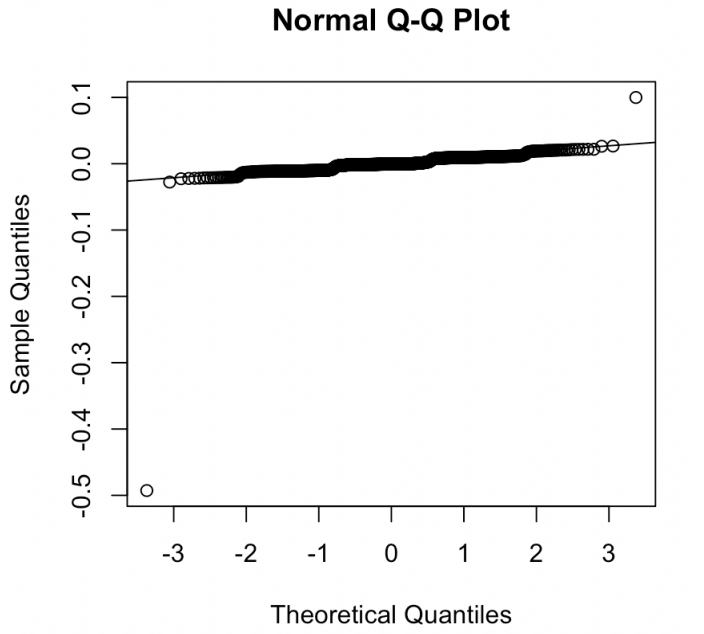
Appendix

Appendix A



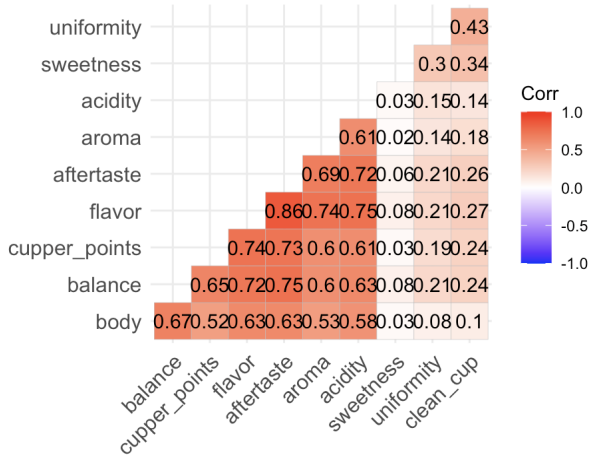
(Scatterplot of Residuals vs Fitted values. The correlation is approximately 0 and most of the points follow the line. We can assume equal variance.)

Appendix B



(QQ plot. Besides the outliers on the tails, the rest of the points are on the line. We can assume it is mostly normal.)

Appendix C



(Correlation Matrix. Coefficients of 0.8 and above are considered highly correlated. From this, we can see Flavor is highly correlated with Aftertaste.)

Appendix D

```
Model 1: total_cup_points ~ aroma + flavor + aftertaste + acidity + body +
    balance + uniformity + clean_cup + sweetness + cupper_points
Model 2: total_cup_points ~ aroma + aftertaste + acidity + body + balance +
    uniformity + clean_cup + sweetness + cupper_points
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1   1327  0.3428
2   1328 28.7401 -1   -28.397 109927 < 2.2e-16 ***
```

(Full vs Reduced Model results. It tests the null hypothesis (H0), where the variables that we removed previously have no significance, against the alternative hypothesis (H1) that those variables are significant. In the above case, p-value is less than 0. We reject the null hypothesis.)

Appendix E

```
Coefficients:
               Estimate Std. Error  t value Pr(>|t|)
(Intercept)   -0.0274738  0.0157638   -1.743   0.0816 .
aroma          1.0005837  0.0021242  471.048   <2e-16 ***
flavor         0.9989316  0.0030129  331.553   <2e-16 ***
aftertaste     1.0023721  0.0027444  365.241   <2e-16 ***
acidity        0.9981821  0.0021816  457.553   <2e-16 ***
body           1.0011854  0.0020511  488.111   <2e-16 ***
balance        1.0025247  0.0020893  479.834   <2e-16 ***
uniformity     1.0028173  0.0010358  968.122   <2e-16 ***
clean_cup      1.0004468  0.0007270 1376.049   <2e-16 ***
sweetness      0.9991390  0.0008631 1157.595   <2e-16 ***
cupper_points  0.9968111  0.0016185  615.871   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01607 on 1327 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 3.736e+06 on 10 and 1327 DF,  p-value: < 2.2e-16
```

(A summary of regression.)

R Code:

```
#install.packages("tidytuesdayR")
#library(tidytuesdayR)
#tt_available()
library(dplyr)

#load the dataset
coffee_ratings <-
readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2020/2020-07-07/coffee
_ratings.csv')

#count bean species
coffee_ratings %>% count(species)

#clean data to include only arabica beans and exclude outliers
coffee_clean <- coffee_ratings %>%
  filter(total_cup_points > 0)

#count bean species
coffee_ratings %>% count(species)

#check data
summary(coffee_clean)

#number of rows
nrow(coffee_clean)
```

```
#count bean species
coffee_clean %>% count(species)

coffee_data <- coffee_clean %>% select(total_cup_points, aroma, flavor, aftertaste, acidity, body, balance,
uniformity, clean_cup, sweetness, cupper_points)
summary(coffee_data)

#mean and sd for descriptive summary
mean(coffee_data$total_cup_points)
sd(coffee_data$total_cup_points)

mean(coffee_data$aroma)
sd(coffee_data$aroma)

mean(coffee_data$flavor)
sd(coffee_data$flavor)

mean(coffee_data$aftertaste)
sd(coffee_data$aftertaste)

mean(coffee_data$acidity)
sd(coffee_data$acidity)

mean(coffee_data$body)
sd(coffee_data$body)

mean(coffee_data$balance)
sd(coffee_data$balance)

mean(coffee_data$uniformity)
sd(coffee_data$uniformity)

mean(coffee_data$clean_cup)
sd(coffee_data$clean_cup)

mean(coffee_data$sweetness)
sd(coffee_data$sweetness)

mean(coffee_data$cupper_points)
sd(coffee_data$cupper_points)

#scatterplot matrix
pairs(total_cup_points ~ aroma + flavor + aftertaste + acidity + body+ balance + uniformity + clean_cup +
sweetness + cupper_points, data = coffee_data)

#fit mlr
```

```
mlr <- lm(total_cup_points ~ aroma + flavor + aftertaste + acidity + body + balance + uniformity + clean_cup +
sweetness + cupper_points, data = coffee_data)

#get resid
res <- mlr$residuals

#histogram of resid
hist(res)

# summary
summary(mlr)

# Plot the residuals
qqnorm(res)
# Plot the Q-Q line
qqline(res)

#constant variance
plot(mlr, which =1)

# Install and load the ggcorrplot package
install.packages("ggcorrplot")
library(ggcorrplot)

# Remove the total cup points column
reduced_data <- subset(coffee_data, select = -total_cup_points)

# Compute correlation at 2 decimal places
corr_matrix = round(cor(reduced_data), 2)

# Compute and show the  result
ggcorrplot(corr_matrix, hc.order = TRUE, type = "lower", lab = TRUE)

#second model
second_model <- lm(total_cup_points ~ aroma + aftertaste + acidity + body + balance + uniformity + clean_cup +
sweetness + cupper_points, data = coffee_data)

summary(second_model)

#compare
anova(mlr,second_model)

library(faraway)
# to use vif() function
round(vif(mlr), 2)

#confint
confint(mlr)
```