# W203 Lab 2 Report: Quality Coffee

Angela Guan, Amy Jung, Jeremy Yeung

## Introduction

Coffee comes in all styles and flavors from its original farm, processing method, roasting time and brewing methods. With approximately one billion people in the world drinking coffee daily[1], it is often an integral part of someone's daily routine. This poses a huge customer base for coffee businesses, which many companies have capitalized on. Simply walking a few blocks down San Francisco or driving a few streets down any town results in passing by multiple coffee shops — from mega-companies like Starbucks to small family-owned businesses. There is competition in the coffee business; therefore, it is crucial for small businesses to distinguish themselves from large chain coffee shops to attract customers. One way to stand out is to offer specialty coffee, often described as the highest grade of coffee available. This type of coffee is defined by the SCAA as "coffee that is free of primary defects, has no quakers, is properly sized and dried, presents in the cup free of faults and taints and has distinctive attributes."[2] In practical terms, this means that the coffee must pass the Coffee Quality Institute (CQI) cupping (grading) system.

To evaluate what components help pass the CQI cupping system, we explore the research question:

***Does the acidity feature of coffee improve Coffee Quality Institute rating?***

Coffee plants are grown all over the world, where the beans are the seeds of the coffee cherry. The beans are then distributed to coffee owners, who clean, triage, and process the coffee. Then, the beans go through a roasting process where they turn from green coffee beans to brown coffee beans[3]. Afterwards, the beans are ground and brewed to make a cup of coffee.

## Data and Research Design

Our data source[4] originates from the Coffee Quality Institute, where each row represents a review of a coffee sample. The dataset contains reviews for 1311 Arabica and 28 Robusta beans originating across many countries and are professionally rated. In addition, data is collected on quality measures, bean metadata, and farm metadata of the coffee beans through its lifecycle. The data only contains information on the coffee bean up until the processing method, where the coffee beans are green. A bean owner can send their coffee sample to be evaluated by the Coffee Quality Institute . Coffee scores close to 80 or higher are eligible for a Q Certificate, which signifies that this coffee is among some of the best in the world and can be deemed "specialty" following the standards of SCA.

We focus on the coffee's acidic profile because acidity is a feature that may be manipulated via different methods in the coffee processing, roasting, and brewing methods. Since we would like to investigate the effect of coffee acidity on rating, we will be predicting the variable total_cup_points, which ranges from 0 to 100. Besides acidity, we expect the features of aroma, aftertaste, bean condition, balance, sweetness, and

[1] https://dealsonhealth.net/coffee-statistics/
[2] https://scanews.coffee/2017/03/17/what-is-specialty-coffee/
[3] https://www.homegrounds.co/does-coffee-go-bad/
[4] https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-07-07/readme.md

processing method of the coffee to be key in determining rating. We will use ordinary least squares regression to understand the effects of these features, with acidity, on rating. The goal is to analyze the magnitude of the coefficient of acidity and how it changes between different regression models.

## Models

To measure quality with the CQI rating score, total_cup_points, we will incorporate the key attributes mentioned above into the models. First we created a baseline model, Model 1, which only includes our key variable acidity. Then we proceed with adding covariates in Model 2, followed by creating an interaction variable to better estimate total_cup_points between acidity and processing method in Model 3.

In our exploratory data analysis, we cleaned our data via an R script, examined missing or null values, and visualized the distribution of our variables of interest. Since the proportion of null or missing values was small relative to the data points we had, we dropped null values. As a result, we still had a large enough sample with over 1000 rows of data. The distribution of processing methods had a majority of "Washed / Wet" methods, and so, we encoded this variable into a binary variable for whether an entry had a processing method of "Washed / Wet" or not. We also examined collinearity using a correlation matrix for all the variables. Below, we see that most of the columns are somewhat correlated to total_cup_points, our outcome variable. Furthermore, we plotted the relationships between pairs of variables, and found that we did not need any drastic transformations.

```
## Warning: The following named parsers don't match the column names: X1
```

```
##                   total_cup_points      aroma aftertaste    acidity       body
## total_cup_points        1.0000000 0.66931783  0.8136102 0.67994729 0.64906558
## aroma                   0.6693178 1.00000000  0.6594509 0.57202780 0.54542467
## aftertaste              0.8136102 0.65945092  1.0000000 0.67080456 0.67224309
## acidity                 0.6799473 0.57202780  0.6708046 1.00000000 0.60576938
## body                    0.6490656 0.54542467  0.6722431 0.60576938 1.00000000
## balance                 0.7642261 0.58432150  0.7560574 0.63141953 0.68504816
## sweetness               0.4522062 0.07435127  0.1353185 0.08080553 0.06776468
##                     balance  sweetness
## total_cup_points 0.7642261 0.45220618
## aroma            0.5843215 0.07435127
## aftertaste       0.7560574 0.13531850
## acidity          0.6314195 0.08080553
## body             0.6850482 0.06776468
## balance          1.0000000 0.11962640
## sweetness        0.1196264 1.00000000
```

## Results

```
# Putting 3 model outputs together
 stargazer(
   model_1, model_2, model_3,
   type = 'text',
   se = list(get_robust_se(model_1), get_robust_se(model_2), get_robust_se(model_3)))
```

```
##
```

```
## ================================================================================================
##                                          Dependent variable:
##                      ------------------------------------------------------------------------
##                                          total_cup_points
##                            (1)                  (2)                    (3)
## ------------------------------------------------------------------------------------------------
## acidity                  5.875***             1.194***               1.160**
##                           (0.415)              (0.308)               (0.527)
##
## aftertaste                                    3.151***               3.202***
##                                               (0.321)               (0.529)
##
## aroma                                         1.390***               1.206**
##                                               (0.248)               (0.588)
##
## body                                           0.321                  0.613
##                                               (0.257)               (0.558)
##
## balance                                       2.069***               2.119***
##                                               (0.313)               (0.511)
##
## proc_wash_wet
##
##
## acidity:proc_wash_wet
##
##
## Constant                 37.818***            21.358***              19.921***
##                           (3.149)              (1.850)               (5.419)
##
## ------------------------------------------------------------------------------------------------
## Observations               1,044                1,044                  310
## R2                         0.462                0.744                 0.695
## Adjusted R2                0.462                0.743                 0.690
## Residual Std. Error   1.934 (df = 1042)    1.337 (df = 1038)     1.461 (df = 304)
## F Statistic           895.986*** (df = 1; 1042) 603.216*** (df = 5; 1038) 138.690*** (df = 5; 304)
## ================================================================================================
## Note:                                                          *p<0.1; **p<0.05; ***p<0.01
```

From the model outputs, we found that acidity is statistically significant in determining the rating of a coffee. Thus it is important for owners of coffee beans to improve acidity if they want to receive higher scores, which will command higher prices. We find that the processing method of coffee beans is not significant on the effect of acidity.

From our Model 3 output, we find that a key quality measure such as aftertaste determines rating by a relatively large margin, followed by balance of the coffee. The appearance or "body" of the bean is the least important in determining rating. Next, we break down the analysis of each model's regression outputs.

**Model 1:** We first build a linear model that includes the key variable of interest, acidity, and the outcome variable. For a 1 unit increase in acidity, we expect total_cup_points to increase by about 5.875. We chose to examine the effect of acidity on our outcome variable because acidity has the highest correlation with total_cup_points. We see that acidity is significant in determining total cup points since the p-value for the coefficient on acidity is less than 0.05.

**Model 2:** Next, we build another model that includes other factors we suspect will influence total cup points, besides acidity. The covariates include aroma, aftertaste, body, and balance. In the results, we see

that the coefficient for acidity decreased drastically to about 1.1935, which indicates that the covariates absorb some of the causal effect. This is probably due to a correlation between some of the additional covariates and acidity, which may lead to a causal effect. The coefficient for acidity is still significant, and this model seems to capture better model fit since total_cup_points can be explained by more than just acidity. In this model, we added these additional covariates to remove omitted variable bias present in Model 1.

**Model 3:** Lastly, we make a model with an interaction term that investigates whether the processing method of the coffee beans affect acidity. By looking at the coefficient, we should be able to tell if the acidity is heterogeneous whether coffee is processed wet or dry. We observed that the coefficient for the acidity term did not fluctuate drastically from about 1.194 to about 1.237, and since the p-value for the acidity coefficient remains less than 0.05, it is significant, indicating acidity does affect total_cup_points.

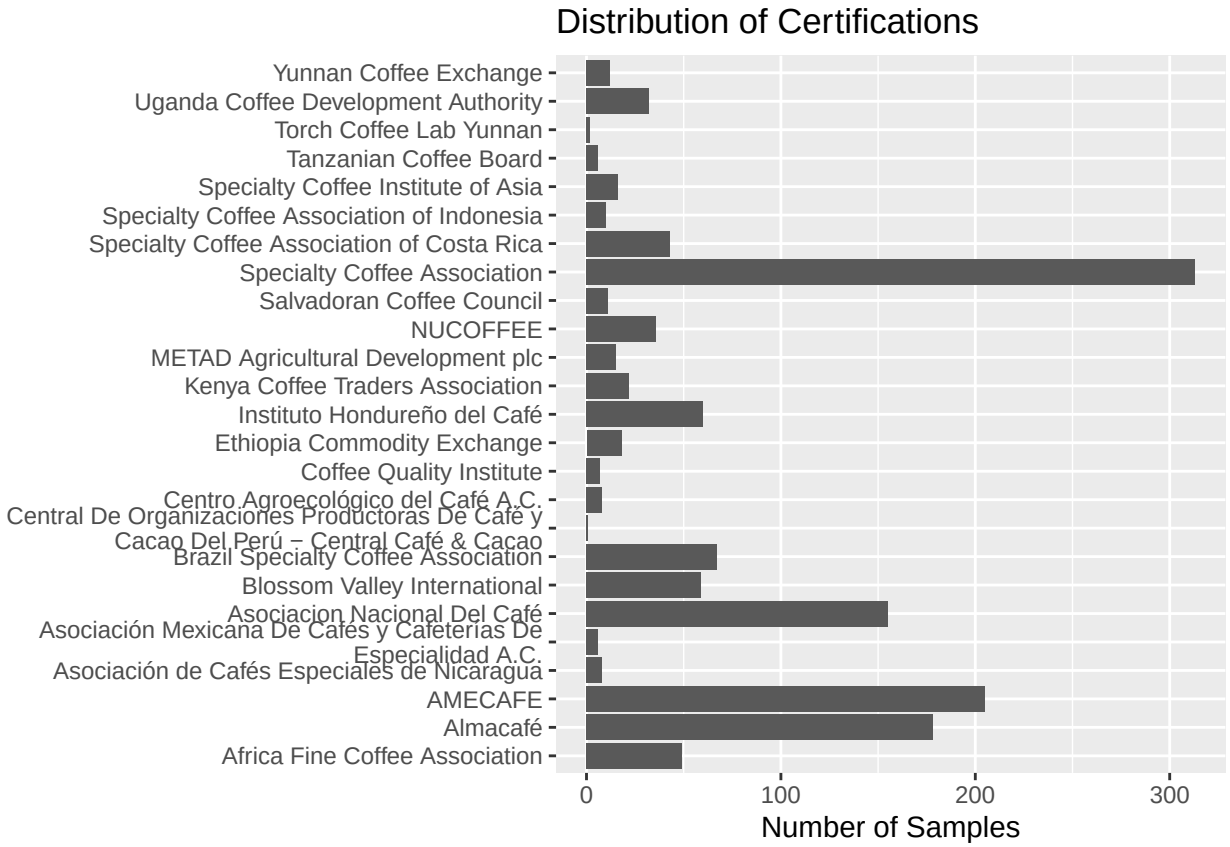
We introduce an additional covariate, processing_method, which represents whether or not the processing method is "Washed / Wet". The processing method is not significant itself. Interacting the processing method with acidity to investigate if the processing method of the coffee beans affects acidity, we find that acidity is heterogeneous whether coffee is processed wet or dry. In other words, the effect of acidity on total_cup_points does not change whether or not the "Washed / Wet" processing method is introduced.

# Limitations

## Statistical Limitations

One of our model assumptions is that our data is independent and identically distributed. Each sample is independent because coffee bean owners send in their coffee sample to be evaluated by the Coffee Quality Institute. Even if a coffee owner submits multiple times, each sample submission is from a different batch of coffee. Since each rating for a coffee bean may come from a different certification, the ratings come from different distributions. As the chart below shows, "Specialty Coffee Association" has the highest number of ratings given. And so, there might be a violation to identically distributed data.

```
ggplot(all_ratings, aes(stringr::str_wrap(certification_body, 50)) ) + geom_bar() + labs(title="Distribu
  ylab('Number of Samples') + coord_flip()
```




## Structural Limitations

Some omitted variables that we were not able to measure and include in the analysis are grind size (granularity of ground coffee), water temperature, and ripeness (time from roast to brewing).

**Grind size:** A smaller grind size results in stronger taste, which increases taste features such as acidity and balance, because there is more surface area that is exposed to the water when brewing the coffee. Although it depends on preference, coffee drinkers typically do not prefer watered down coffee; therefore, grind size is negatively correlated with total_cup_points. And so, there is a positive omitted variable bias (OVB) that points away from 0, so we measure an effect on acidity which is larger than what it actually is due to OVB.

**Water temperature:** As long as the water temperature is in the optimal coffee brewing range, higher temperature water extracts coffee more efficiently, which results in more flavor. So there is positive omitted variable bias present.

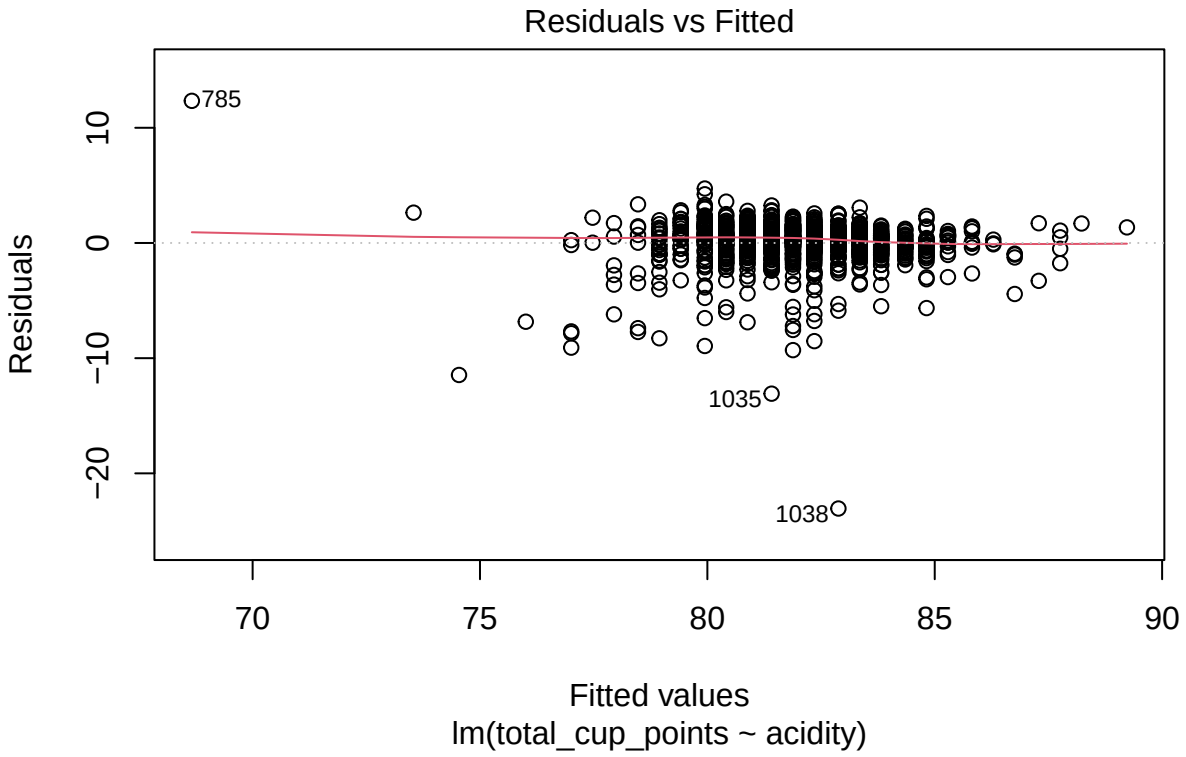
**Ripeness:** After brewing, a period of degassing is required. As the ripeness, or time between roast and brew, increases, more gas is allowed to depart the mean, which makes it more permeable for water when it comes to extracting. If water can't permeate suitably during coffee extraction, this may result in a weak and sour flavor profile. Similarly, this would be positive omitted variable bias.

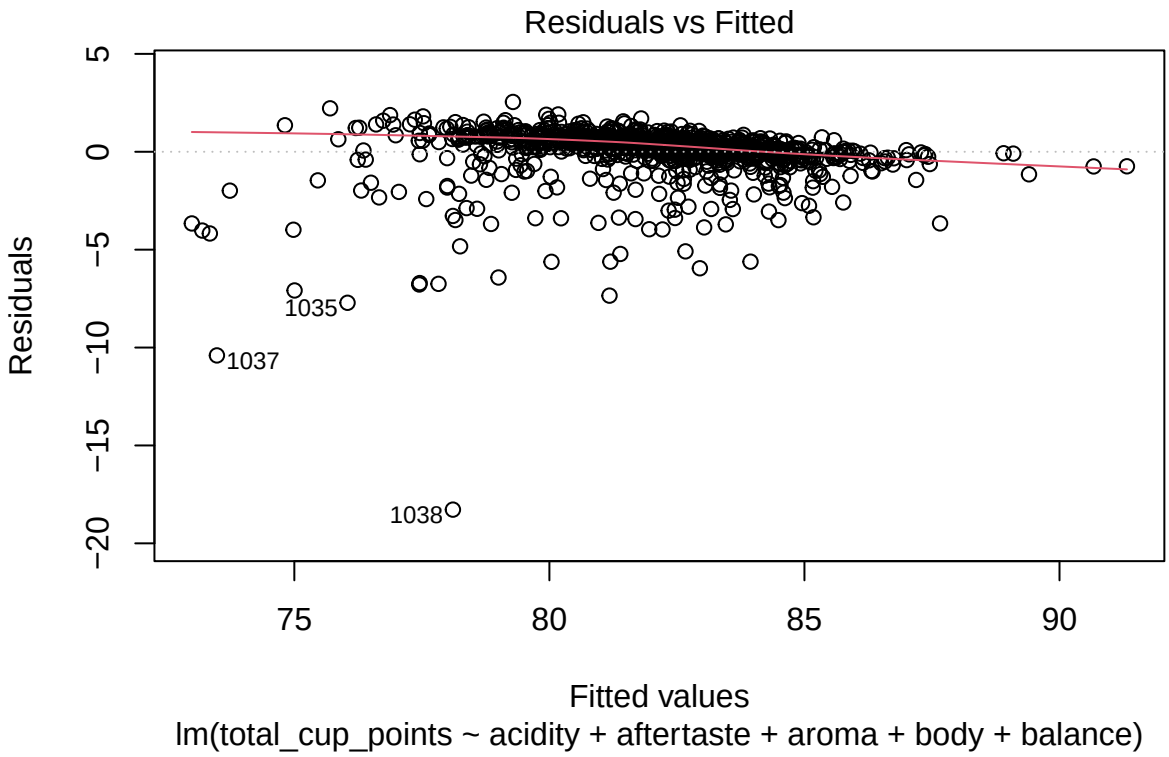
The omission of this variable does call into question the core results. However, our data comes from the Coffee Quality Institute, where we assume there is a standardized process for evaluating coffee quality. Therefore, we assumed the extraction and brewing methods are consistent across all coffee. Future data we could collect that would resolve any omitted variables bias may include coffee extraction in brewing data.

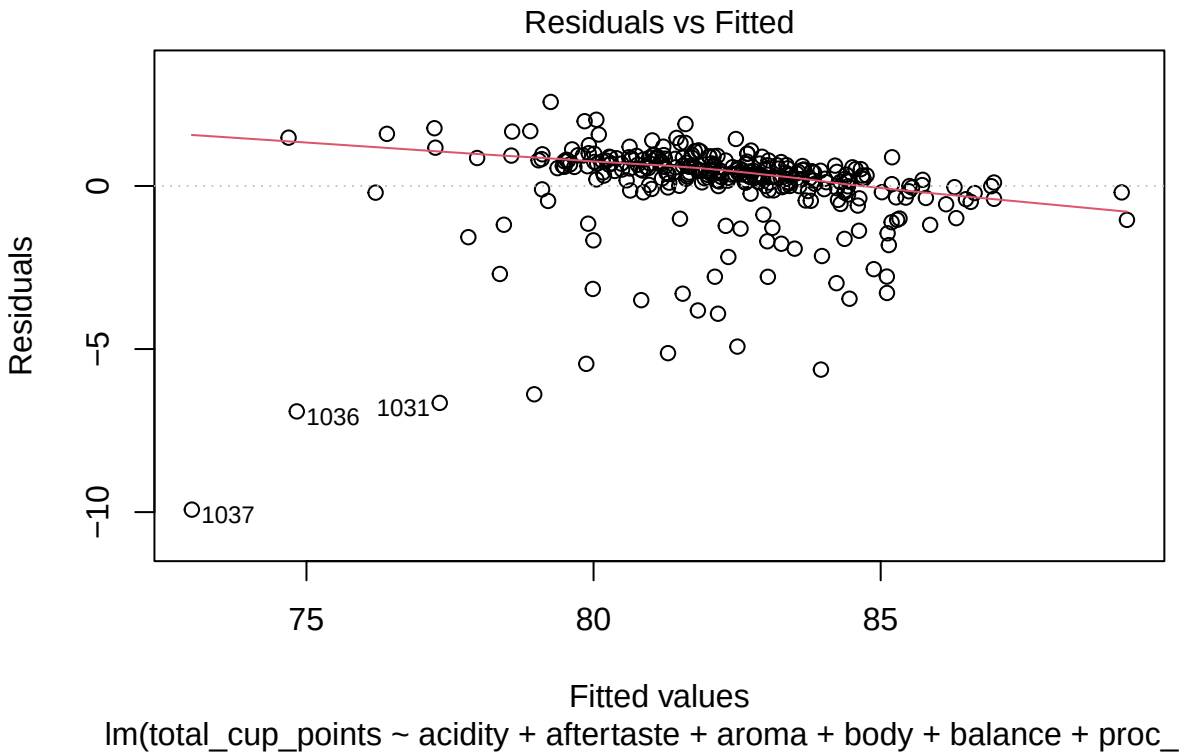
## Other CLM Assumptions

**No perfect colinearity:** Running our model lm(y ~ x), did not drop any variables, which indicates no perfect colinearity. In addition, our correlation matrix shows that none of the variables are perfectly colinear (ie. correlation equal to 1).

**Linear conditional expectation:** Upon looking at the residual vs. fitted values, model_1 and model_2 have a flat line (zero slope) of conditional expectation, therefore, upholding the linear conditional expectation assumption. However, model_3 shows a slight downward negative slope for the residual vs. fitted values, therefore violating this assumption. In this case, we should consider non-linear relationships when modeling model_3, like a polynomial relationship, by transforming the response and/or predictor variables.

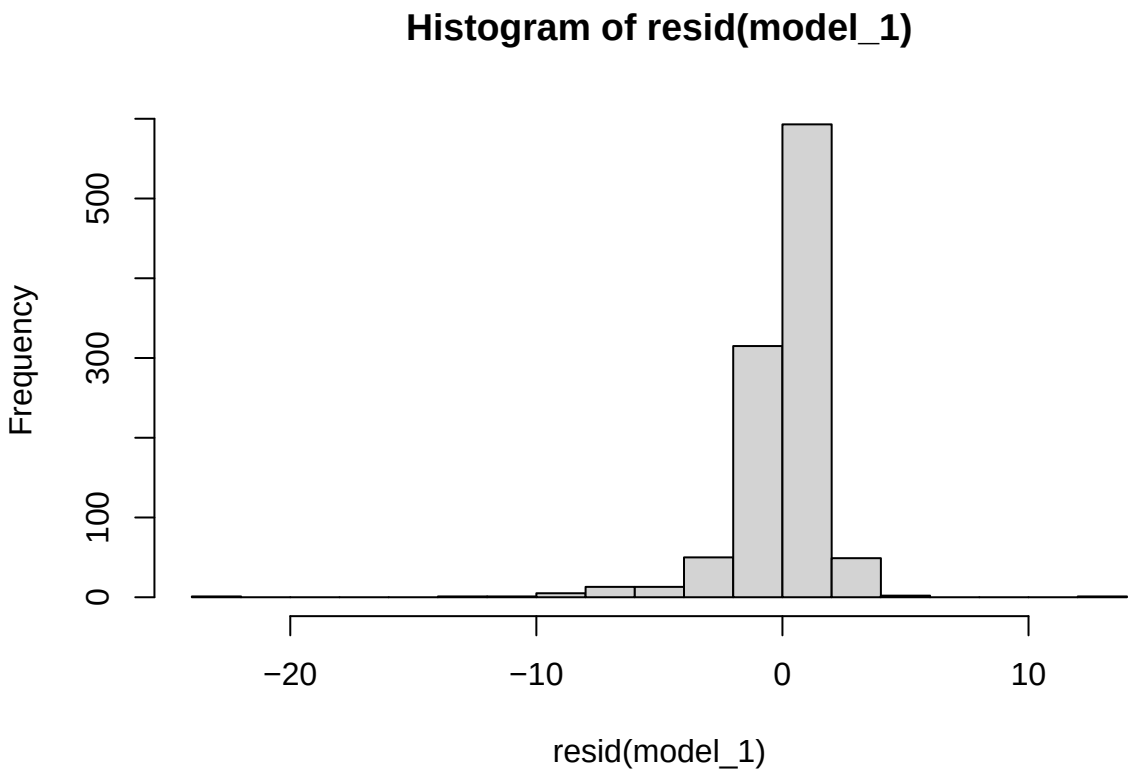


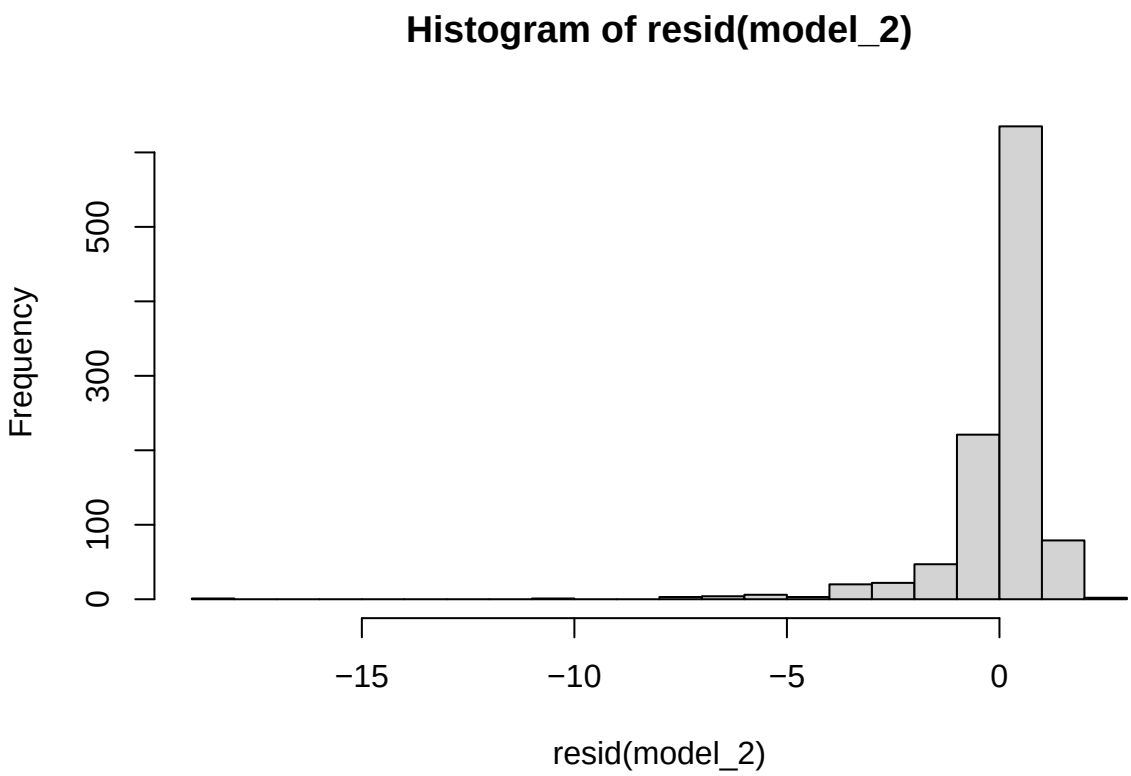

Residuals vs Fitted
Residuals
5
0
−5
−10
−15
−20
1035
1037
1038
75
80
85
90
Fitted values
lm(total_cup_points ~ acidity + aftertaste + aroma + body + balance)

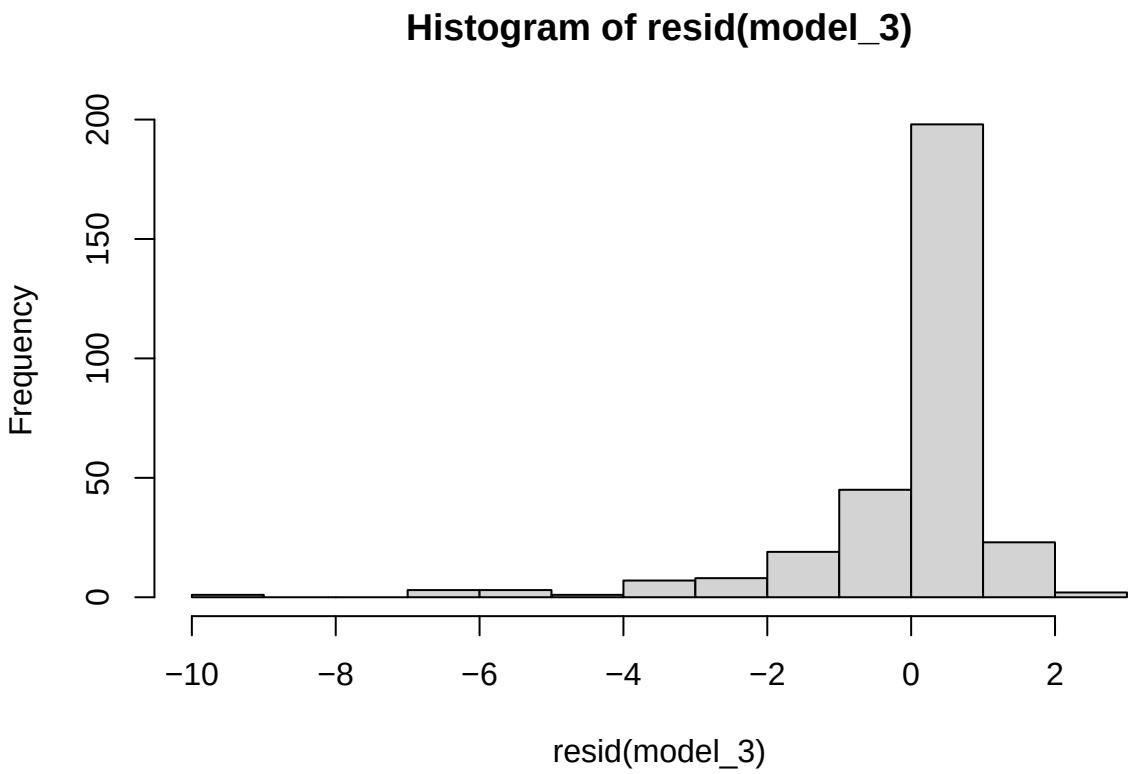



**Homoskedastic errors:** The residual vs. fitted graphs above show very little change of band thickness (how fat/thick the scatter is). Therefore, we conclude that the homoskedastic errors assumption is met for all three models.

**Normally distributed errors:** As seen below, the residuals for all three models look a bit skewed (not completely normally distributed); therefore, the assumption of normally distributed errors may not stand. In this case, we can use the Box-Cox (1964) method for choosing the best transformation from the set of power transformations to correct for this violation.


Histogram of resid(model_1)
Frequency
0
100
300
500
−20
−10
0
10
resid(model_1)


Histogram of resid(model_2)
Frequency
0
100
300
500
−15
−10
−5
0
resid(model_2)




## Conclusion

In conclusion, we found that acidity does have a significant effect on determining CQI cup points. This was supported by the results in our first regression model which only included acidity as the covariate. In our second regression model which included multiple covariates, we found that acidity was correlated to attributes such as aftertaste and aroma, which reduced the effect of acidity on cup points. Our last regression model with the interaction term between processing method and acidity showed that the effect of acidity on total_cup_points does not change whether or not the "Washed / Wet" processing method is introduced.

Based on the results, we recommend that further research is conducted to look at different aspects that go into producing a cup of coffee. For example, since our current data set only contains information on the processing step of coffee production, data on the roasting and brewing methods may give a more complete model on factors affecting coffee quality rating. Factors listed in our "Structural Limitations", like grind size and water temperature, may provide key insights into manipulation of coffee attributes, such as acidity, and ultimately coffee rating.