Regression Lab: What makes coffee good?

Adam Hyman, Anstonia Ma, Emily Huang

April 11, 2022

Contents

1 Introduction

Starbucks is one of the most recognizable brands of coffee in the world, but recent surveys conducted by leadership has shown a negative attitude towards the quality of our drinks; one of the most notable criticisms was “that the coffee tastes bad. The processes used are seen as clearly inferior to anyone who knows the first thing about coffee. Or anyone who has tried a straight espresso from one of their branches. But at the same time it is an incredibly successful franchise” from this source. Because of these criticisms, as part of the Starbucks Data Science team, we’ve been tasked to see whether or not taste is a strong indication of good quality coffee. We want to learn if our coffee beans is the reason for causing poor feedback on the coffee quality in hopes of changing public sentiment of our products. To be clear, our research question for this analysis is “Is coffee flavor a significant indicator of coffee quality?” with our null hypothesis being “flavor has no significant effect on coffee ‘quality’ ”.

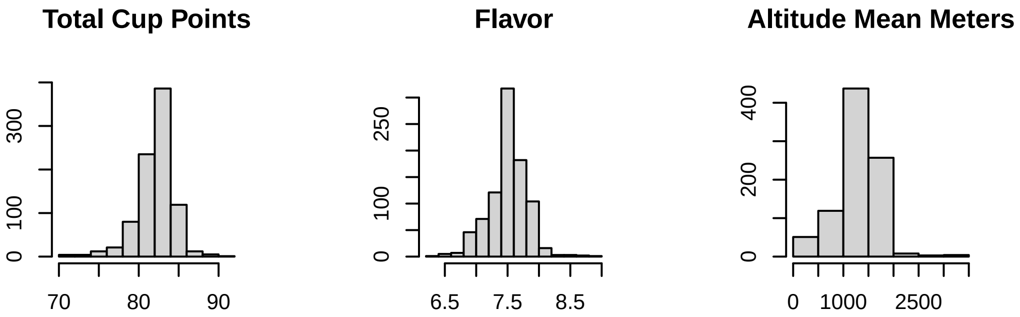
2 Data and Research Design and Model Building Process

In order to better understand what traits of a coffee bean indicate good quality, we’ll be using data collected from the Coffee Quality Institute’s review pages in January 2018 where all of the coffee are rated by certified coffee drinkers, which contains 1339 observations of different coffees and their attributes like aroma, flavor, body, altitude, uniformity, country of origin, and etc. Because Starbucks coffee beans are only of the Arabic species and from Latin America, Africa, and Asia-Pacific regions, we want to filter our data to best fit these sourcing conditions. We also want to ensure that the panelists tested a large batch of coffee rather than a small subset to ensure the scoring was based on a good sample of coffee beans, which meant we filtered out coffee bean observations that had less than 10 bags included in the judging.

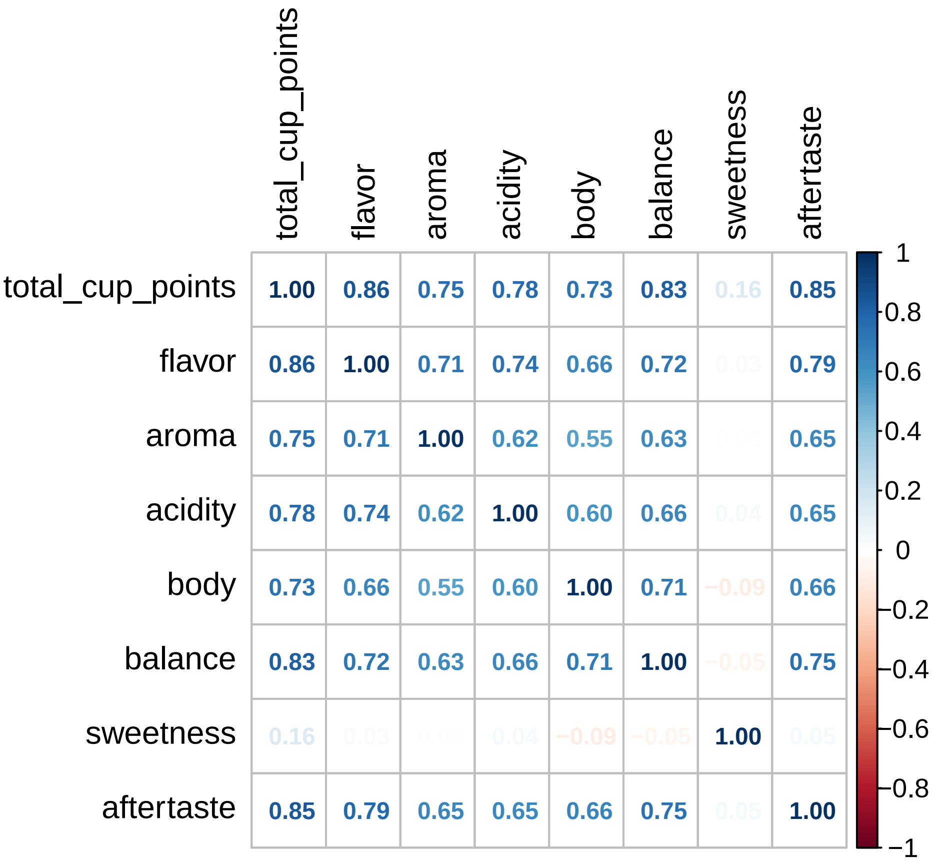
Cause	Number of samples available for analysis (after removal for cause)	Remove number of samples for cause
Start	1339	
Arabica species coffee	1311	28
Coffees with relevant countries of origin for Starbucks	1225	86
Coffees with sample bags >= 10	1079	146
Coffee with reasonable altitude mean meters	1075	4
Coffees with valid values for our outcome, treatment, and predictor variables	886	189
Coffee with greater than 70 “total cup points”	879	7

We additionally did some data cleanup by filtering the “harvest year” to be purely the year since there were varying formats in the column such as a specific date, a time period with only months, and other variations. Because we were only interested in just the year, we filtered out any observations that did not have the year in the field at all and modified the rest of the fields to only have the year. Next, we filtered for our control variable “certification body” such that all subsidiaries of “Specialty Coffee Association” were under the same body. Lastly, we filtered out mean altitudes where the value was greater than 10,000 since it didn’t make sense logically for coffee beans to be grown at such a high altitude (for context, Mount Everest is 8,894 meters above sea level). We’ll be creating 3 separate models to achieve separate goals for our analysis. The first model will include just our key treatment and outcome variables. The second model will include all variables from the first model in addition to three control variables in order to ensure that the control variables do not absorb the causal effect of flavor. The third model includes all variables from the second model with the addition of variables that we believe directly impact flavor to lessen omitted variable bias from the second model as these variables can have a direct impact on flavor and "total cup points."

We identified our response variable to be "total cup points" which is the grader's "holistically integrated rating of the sample as perceived by the individual panelist" (and therefore a metric value as graders have extensive training to be accurate by the decimal) to help define coffee quality. Our treatment variable is the coffee "flavor" which is also a score determined by the Coffee Quality Institute graders and is a metric variable. We also identified "country of origin", "certification body", and "altitude mean meters" to be our control variables to ensure those variables were accounted for when testing how much "flavor" affected the "total cup points". For our control variables, there are 25 countries ("country of origin"), 19 certification bodies ("certification body"), and the "altitude mean meters" which is the average height where the coffee is grown. While we decided to use these as control variables to ensure that flavor was a significant variable regardless of bean origin and grading body, there could be some problems with the control variables we chose. For instance, certain countries and certification bodies are over indexed in our sample and that certification bodies have a relationship with the country of origin because of the location of said certification body. However, we believe that there are enough graders within each respective certification body to be independent and our country sample to be robust enough to overcome these problems. We did not feel a need to transform the variables because the metric variables that we put into our model were approximately normal and our other variables (such as certification body) were categorical in nature.



Our first model is with our response variable ("total cup points") and treatment ("flavor"). The second model will include the variables in our first model with the addition of three control variables ("country of origin", "certification body", and "altitude mean meters").



The third model's additional variables were chosen from the above correlation matrix where we saw that "aroma", "acidity", "body", "balance", and "aftertaste" are correlated with "flavor" and "total cup points." While the correlation matrix above does not show as heavy correlation as "body" or "acidity", we also included "sweetness" as "sweetness" directly impacts the flavor of the coffee. However, we did not include "clean cup" or "uniformity" because of their weak correlation in our matrix, their widely skewed data that we were unable to normalize with transformations, and also weak practical significance to "total cup points" and "flavor". We also chose to include "harvest year" as an additional variable as we believe it would be important to include it as the year that the coffee is grown (even though the coffee is graded within a year of harvest) could have an effect as well but not as a control variable in our second model. If flavor is consistently statistically significant throughout all three of these models, then we can confidently conclude that flavor does have a significant effect on "total cup points" which in turn means that flavor is a significant indicator of coffee quality as we are using "total cup points" as a measure of coffee quality.

To verify that these models are different and also perform differently from each other, we will be performing 2 anova tests to verify that the variables we added to each are statistically significant.

```
## Analysis of Variance Table
##
## Model 1: total_cup_points ~ flavor
## Model 2: total_cup_points ~ flavor + country_of_origin + certification_body +
##     altitude_mean_meters
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    877 1277.5
## 2    834 1061.0 43    216.53 3.9582 3.764e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Model 1: total_cup_points ~ flavor + country_of_origin + certification_body +
##     altitude_mean_meters
## Model 2: total_cup_points ~ flavor + country_of_origin + certification_body +
##     altitude_mean_meters + harvest_year + processing_method +
##     aroma + acidity + body + balance + sweetness + aftertaste
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    834 1061.00
## 2    817  551.44 17    509.56 44.408 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we see above, all of our anova tests all lead us to reject the null hypothesis and assume that the variables we used are statistically significant. While we have shown that they are significantly different, we also want to verify that the performance of these models are also different. To do this, we will be calculating the mean squared error for all of our models and comparing how each of our models stack up against each other.

```
## [1] "Model 1 MSR: 1.4533870262719"
```

```
## [1] "Model 2 MSR: 1.20705439808781"
```

```
## [1] "Model 3 MSR: 0.627353046333709"
```

With each progressive model, our MSR improves with our greatest improvement being from our second to our third model. This aligns with the goal of our third model since we included all the variables that we believed would have an impact on “flavor” and “total cup points” so their inclusion should also help our model’s predictive power.

3 Results

```
##
## ================================================================================================
##                                                 Dependent variable:
##                     ----------------------------------------------------------------------------
##                                                   total_cup_points
##                               first                    second                  third
##                                (1)                      (2)                     (3)
## ------------------------------------------------------------------------------------------------
## flavor                       (0.135)                   (0.147)                 (0.215)
##                              6.378***                  6.048***                1.970***
##
## aroma                                                                          (0.149)
##                                                                                1.019***
##
## acidity                                                                        (0.154)
##                                                                                0.850***
##
## body                                                                           (0.167)
##                                                                                0.476**
##
## balance                                                                        (0.153)
##                                                                                1.602***
```

```
##
## sweetness                                                                       (0.112)
##                                                                                 1.512***
##
## aftertaste                                                                      (0.198)
##                                                                                 1.732***
##
## Constant                     (1.012)                   (1.808)                  (1.864)
##                             34.259***                 36.247***                 9.415***
##
## ------------------------------------------------------------------------------------------------
## Observations                   879                       879                      879
## R2                            0.719                     0.767                    0.879
## Adjusted R2                   0.719                     0.754                    0.870
## Residual Std. Error      1.207 (df = 877)          1.128 (df = 834)         0.822 (df = 817)
## F Statistic         2,244.535*** (df = 1; 877) 62.279*** (df = 44; 834) 97.048*** (df = 61; 817)
## ================================================================================================
## Note:                                                              *p<0.05; **p<0.01; ***p<0.001
```

Stargazer shows that the r-squared improves as more variables are added to the model, when we move from the first model to the second, as well as from the second to the third. As can be seen in the chart above, all numeric explanatory variables have p-values below 0.01, which shows they are very significant. Categorical variables, such as country of origin, were also also statistically significant, but were omitted from the chart for readability.

With the results of the test, we've discovered that flavor is such a strong indicator of coffee quality that it explains 72% of the variation of total cup points. This means that we need to heavily consider customer feedback regarding the coffee taste; if coffee taste is as bad as customers are saying in recent surveys, we need to investigate what has affected the coffee beans to have caused poor quality coffee. There are other factors that we need to consider regarding customer feedback such as the types of drinks customers complain about poor taste. If customers are complaining about drinks that aren't purely just coffee, there may be additional surveying and data analysis we need to consider before making a definitive statement about what is causing the poor taste in our coffee. More details about what we should gather for data will be detailed in the "Structural Limitations of the Models" section. At the very least, we know our customers are giving an indication of our products' quality that requires high attention in fixing as soon as possible.

4 Limitations of the Models

4.1 Statistical Limitations of the Models

The large sample assumptions to be evaluated are:

1. **Independent and Identically Distributed (I.I.D.) Data:**

Our initial dataset contains all coffees featured in the Coffee Quality Institute's review pages in January 2018. The featured coffees are those selected by experts that are of interest to serious coffee drinkers. The experts intentionally selected a variety of coffees, based on characteristics like uniformity, body, aroma, flavor, etc to ensure that a wide range of coffees were featured, which mitigates our concerns about the dataset. Coffee submitted for grading is from all over the world so that the data won't be skewed toward just one region and graders have a standardized approach to analyzing the coffee and rating it equally regardless of origin. We were careful that the records that we chose to exclude did not cause this assumption to be violated.

2. **A Unique BLP Exists:**

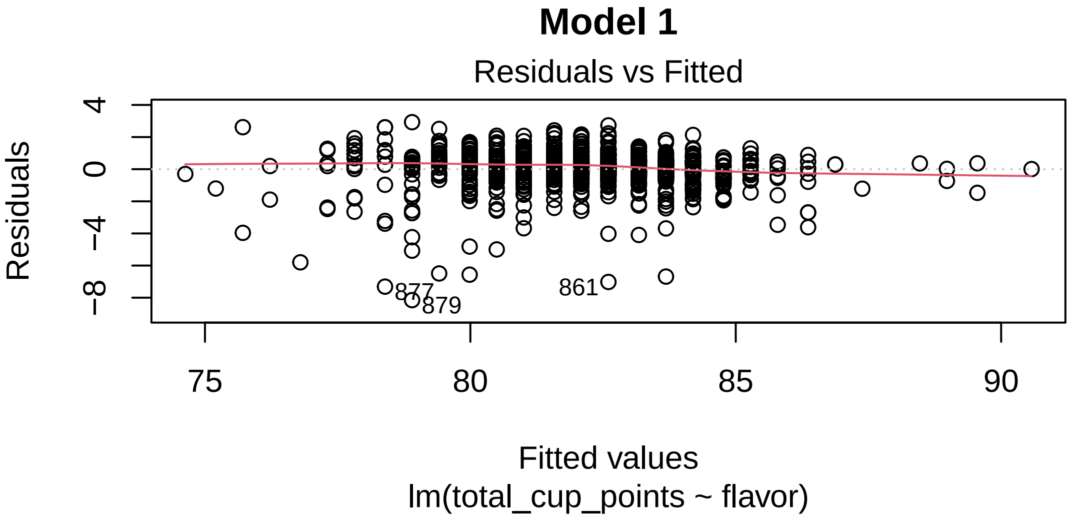
We verified that a unique BLP exists by ensuring that there is no perfect collinearity. We tested for collinearity using a correlation matrix above in the Model Building Portion.

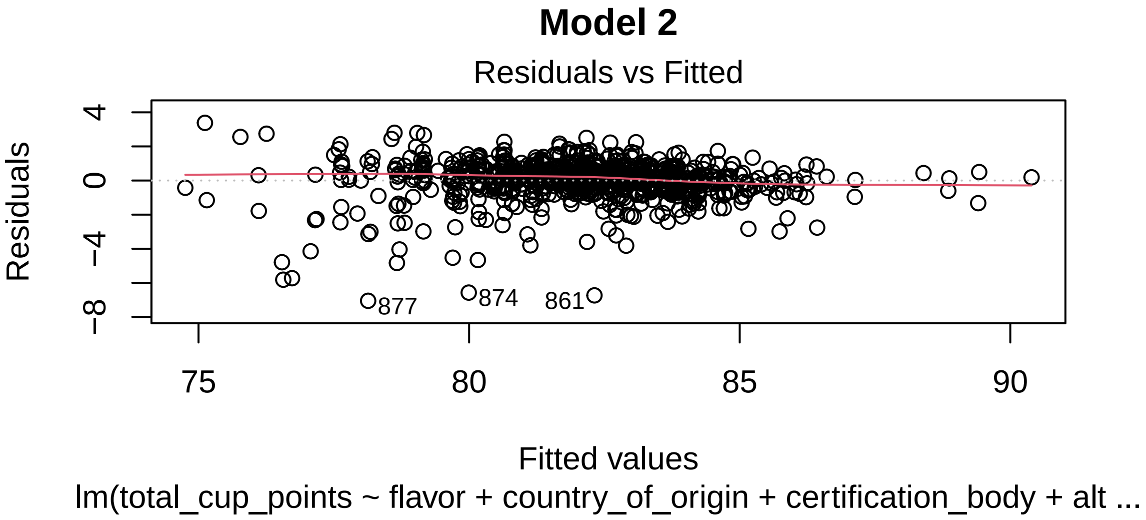
As none of our variables had a coefficient of 1.0 besides when they were compared to themselves, our correlation matrix does not show perfect collinearity between any of our predictors.

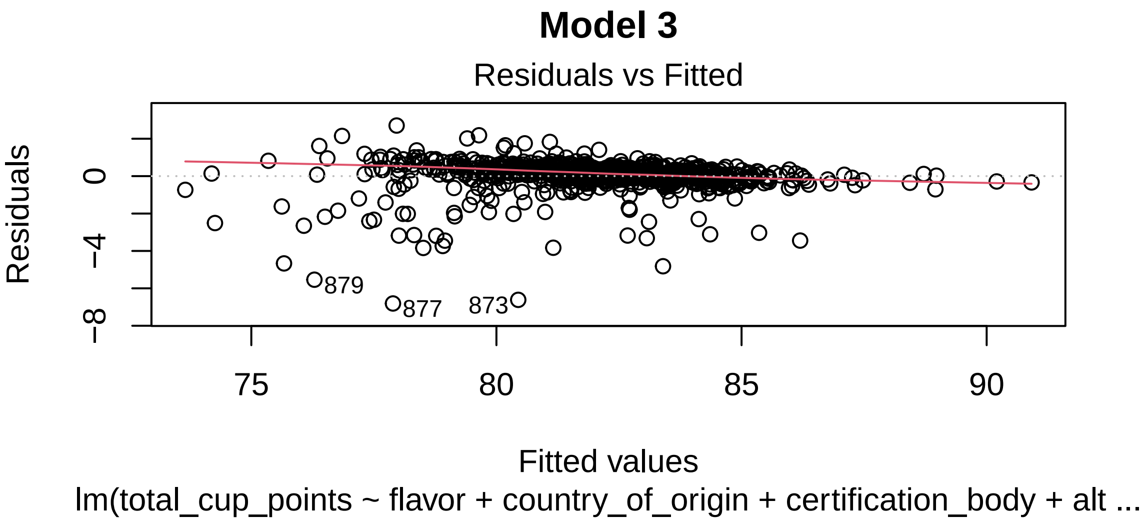
3. **Linear Conditional Expectation:**

We validated that there is a linear relationship between predictors and the target variable by looking at the residuals and checking whether they are consistent for various inputs.

We see that this is the case below for all three of our models, where the residuals are equally dispersed for different fitted values. This satisfies the assumption of linear conditional expectation.

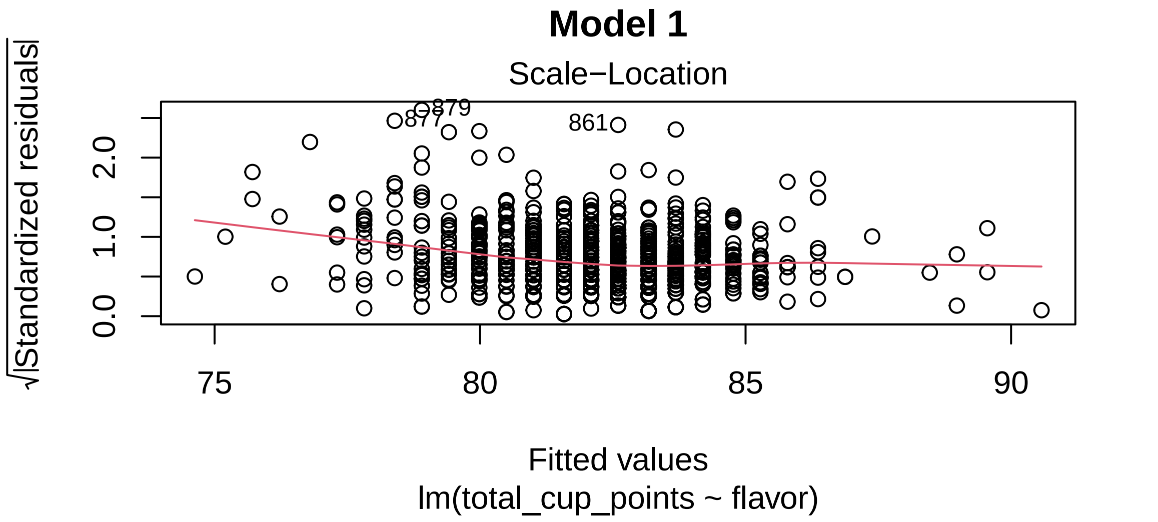


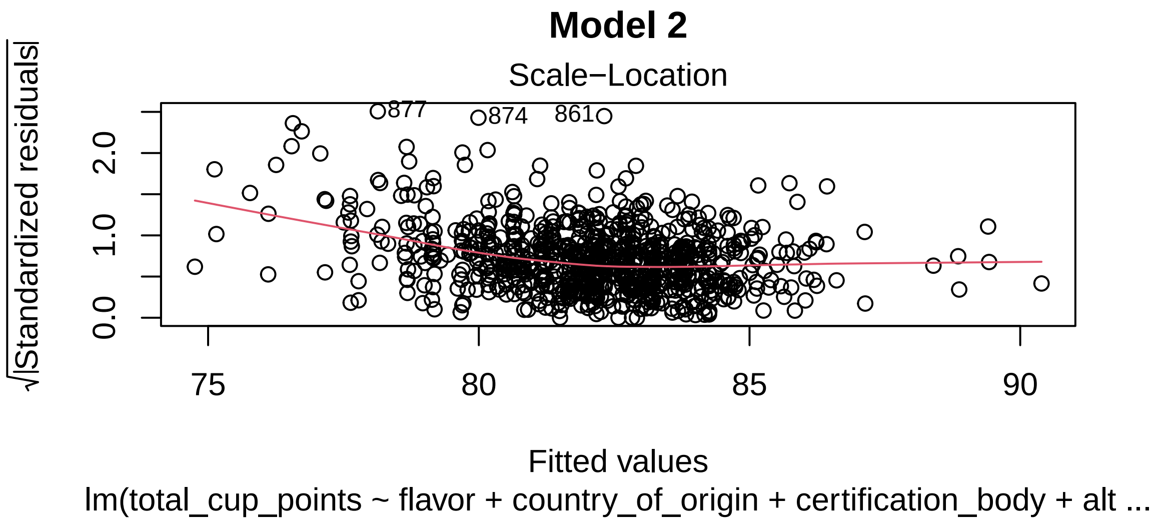


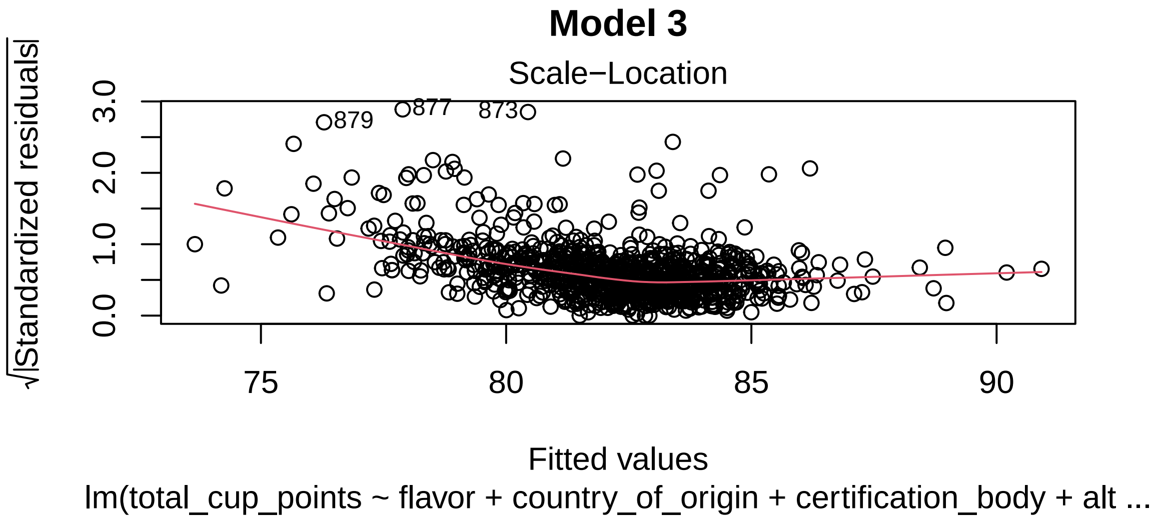


4. **Homoskedastic Errors:**

We are checking if the square root of the residuals are equal across the regression line. For our first model, the line is roughly horizontal across the graph so we believe this satisfies the assumption of homoskedastic errors. However, we do see for our second and third models that our residuals show a steep decline from the left side of the chart to the middle where our points are congregated. What this would mean is that for coffees that are ranked lower than the average, there is more variation in our reesults. Once the coffee rating (total cup points) increases as we move to the right of the graph, the line becomes much more horizontal to showcase equal variance of our residuals. Following this reasoning, while the line is not the ideal horizontal that we would prefer for our models, because the skew on the left can be explained by the lower amount of "bad" coffee which can vary in terms of why the coffee quality was poor while "good" coffee tends to share similar traits and are more consistent. Hence, we say that all of our models satisfy this assumption.

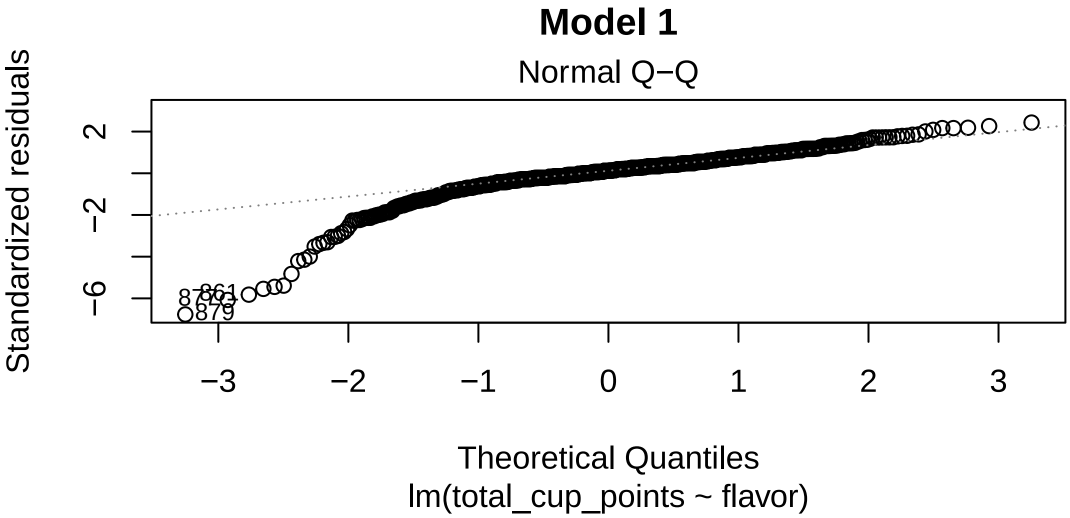


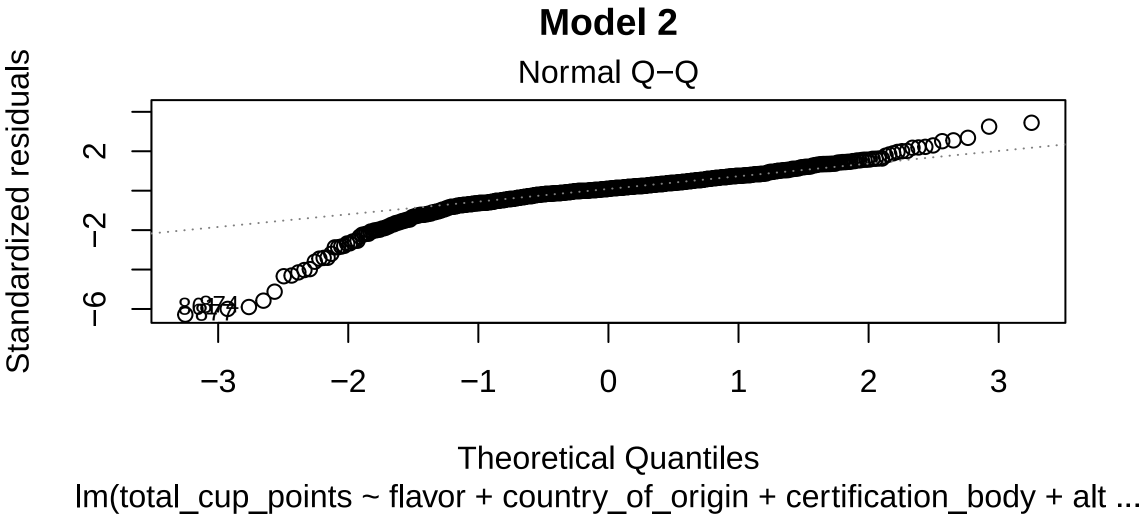


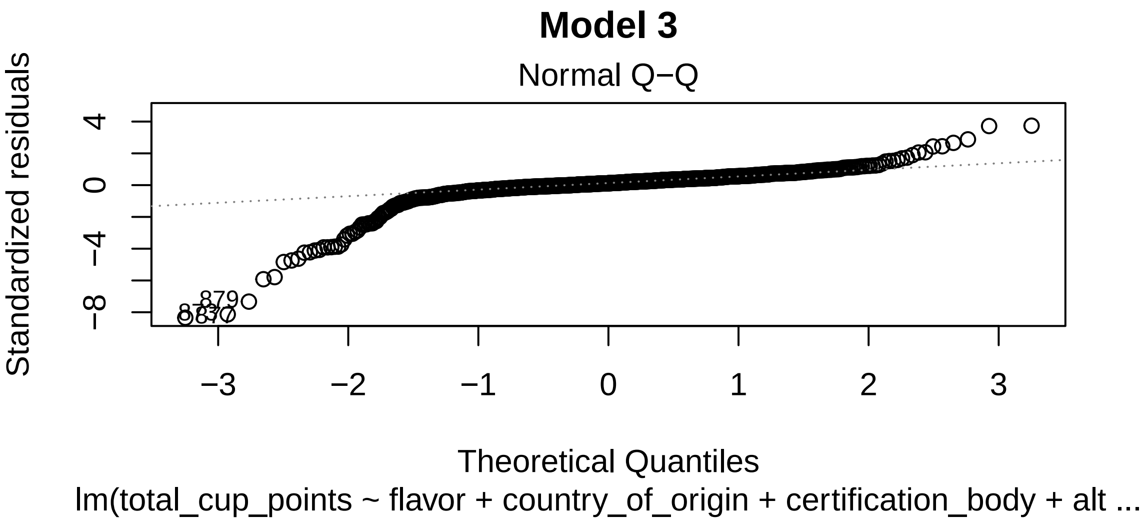


5. **Normally Distributed Errors:**

We used Q-Q plotd to check that variables are multivariate normal. The Q-Q plots below show that the data is very close to normal, however it has slightly fatter tails, which is why the Q-Q plots are slightly below the line on the far left and slightly above on the far right. However, it's close enough that we consider that this satisfies the assumption of Normally Distributed Errors.


Model 1
Normal Q−Q
Standardized residuals
2
−2
−6
861
879
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(total_cup_points ~ flavor)


Model 2
Normal Q−Q
Standardized residuals
2
−2
−6
874
877
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(total_cup_points ~ flavor + country_of_origin + certification_body + alt ...



4.2 Structural Limitations of your Model

“Uniformity” and “clean cup” were omitted because they had a low correlation with “total cup points”. They could have had an effect on our final model.

We think that “uniformity” and “clean cup” would have been positively correlated with “flavor” (a predictor) and “total cup points” (the response variable). Because those correlations are both positive and the correlation coefficient of “flavor” is also positive, the omission results in positive omitted variable bias which increases the coefficient of “flavor”, pushing it further away from zero.

A further limitation of our model is omitted variables which impact the final product that is consumed by the consumer like machie that is used to grind the coffee (categorical variable). There are different type of machines (hypothetically split into “good quality” and “bad quality” machines) which will have different effects on both flavor and “total cup points.” We posit that a good quality machine will have a positive correlation with flavor and also a positive correlation with “total cup points.” This will lead us to a positive omitted variable bias and increase the statistical significance of “flavor.” However, inversely, a poor quality machine will have a negative correlation with flavor and also a negative correlation with “total cup points” which will still result in a positive omitted variable bias.

Omission of these variables does not call into question the core results, but should lead to further study as coffee is not always drunk by itself, but is combined with other additives like sugar, milk, etc.

5 Conclusion

The results of our analysis found that the taste of a coffee is actually a strong indicator of coffee quality which means our customers may be indicating something significant about our products. From the previous sections, we know there is more additional data collecting and analysis that needs to be done in order to pinpoint what is making customers have negative sentiment towards our products.

We should work with Quality Control to see if there is anything wrong with our coffee beans before moving on to potential next steps and seeing if there is other factors coming into play with our drink quality. Those next steps could be contacting the Coffee Quality Institute to help find out more about the testing method they used to give cup points, surveying customers about their favorite drinks, least favorite drinks, and their reasoning behind their picks, and figuring out if our equipment may be affecting the drink quality rather than our ingredients. Having data collected from one or more of these potential steps may help give light on what our coffee needs in order for critics of our products to view us more favorably.