Estimating Coffee Ratings - Dataset from Coffee Quality Institute

Abraham Yang, Erica Nakabayashi, Melissa Olivera

2023-04-21

Limitations

To assess our model limitations, we considered both statistical and structural limitations. For the statistical limitations there are 2 large-sample statistical assumptions we evaluated. We first evaluated whether the data was IID. We concluded that this assumption was not met given that the data collected primarily focused on top coffees. >85% of the reviews had a total cup score >80 pts. We believe this is not a true representation of the population and geographical clustering is also a factor.

The second statistical assumption we evaluated was unique BLP/no perfect collinearity. For this assumption, we evaluated the coefficients of each of our models and noticed that R did not drop any of our variables, indicating no perfect collinearity. This assumption also includes the requirement that a unique BLP exists, however our distributions had heavy tails so this assumption was not met.

For structural limitations, we identified Omitted Variable bias and Right Hand Side bias. For omitted variable bias, there were several variables we omitted that may bias our results such as moisture and variety due to too many missing values. These variables are positively correlated to total cup points, so we would expect a bias moving away from zero and thus making our hypothesis tests overconfident.

For RHS bias, we took this into consideration in our initial model development. One of the initial models we considered was using country of origin and altitude to predict total cup points. We soon realized that this would result in altitude as a RHS variable, as country can impact altitude and both altitude and country can impact total cup points. To overcome this, we instead used subregion in our models.

Conclusion

Tying this back to the original research question "How do geographical features such as altitude, climate and regional location affect coffee cup scores?", we determined that our analysis was inconclusive and we are not confident that these features affect coffee cup scores.

Our top performing model (using altitude, subregion and tropical to determine cup points) suggests that there is a positive relationship between cup points and altitude, a positive relationship between total cup points and subregions Eastern Africa, Eastern Asia, and South America, and a negative relationship between total cup points and being partially tropical. However, even though this model had good statistical significance, we ultimately decided it is unreliable.

There are many other factors we did not consider in this OLS regression model largely due to missing or inconsistent data. If we were to move forward with this research question, we would collect new data which accurately represents the coffee population and we would ensure data was complete and consistent across all reviewers