Multiple Linear Regression:

What makes a good cup of coffee

Presented by: Jeanette Nguyen

Agenda

- Questions and Goals
- Data, Method, Summary statistics
- Multiple Linear Regression model
- Assumptions
- Results
- Conclusion

Questions and Goals

What makes a good cup of coffee?

Estimate the association between the response and explanatory variables.

From the explanatory variables, what is most and least important?

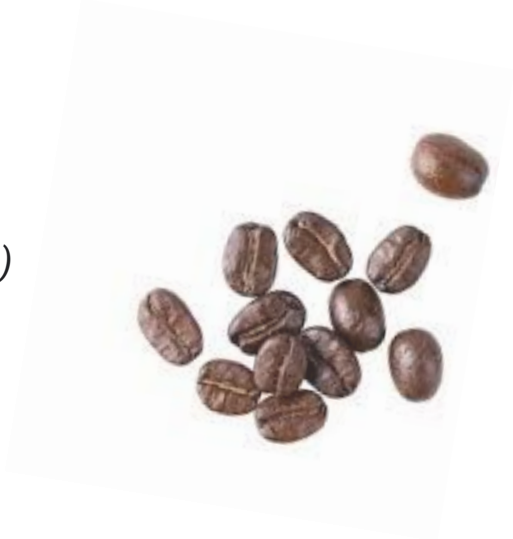
<u>Hypothesis</u>

H_0: $\beta_1 = \beta_2 = \ldots \beta_p = 0$ (p=10)

(There is no useful linear relationship between y and any of the p predictors)

H_A: at least one β_j does not equal 0 (i=1,...,p)

(At least one β is not 0, the model is deemed useful)



Data, Method

Method

Data is collected from the Coffee Quality Institute's review pages in January 2018 of reviews of 1338 Arabica and Robusta coffee beans. 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.

Variables

Response = Coffee Ratings (Total Cup Points); rated from 0-100

Explanatory = Aroma, Flavor, Aftertaste, Acidity, Body, Balance Uniformity, Cup Cleanliness, Sweetness, Cupper Points; rated from 0-10

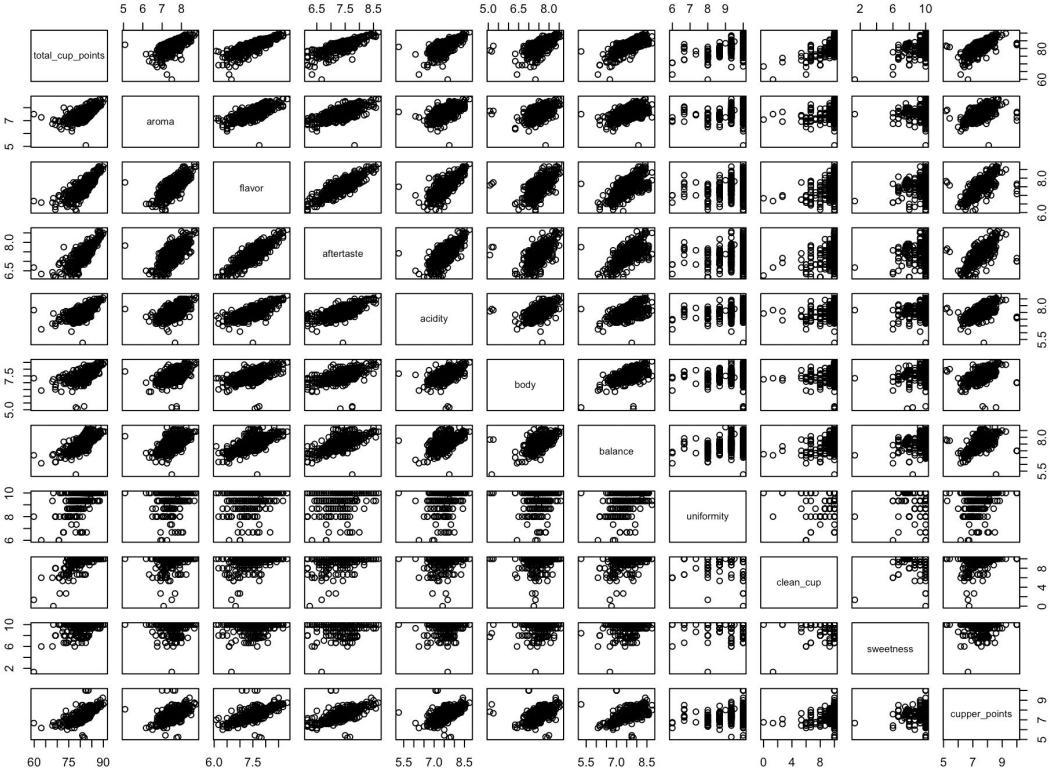
Multiple Linear Regression model

$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \ldots + \beta nXn$

Total Cupping Points = β_0 + β_1 aroma + β_2 flavor + β_3 aftertaste + β_4 acidity + β_5 body + β_6 balance + β_7 uniformity + β_8 cup cleanliness + β_9 sweetness + β_{10} cupper points + ε

- β_0 : the average total cupping points for coffee beans with variables units = 0
- $\beta_{1\ldots}$ β_{10} : the average difference in total cupping points for coffee beans whose predictor variable differs by one unit.
- ε : the model error residuals

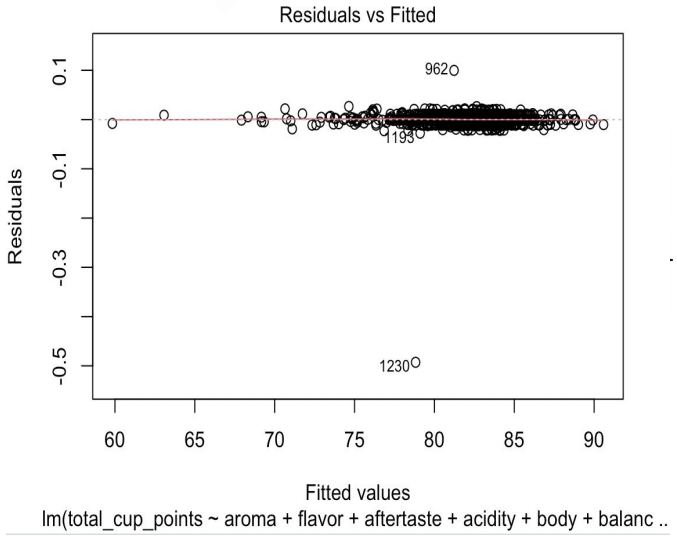
Summary Statistics

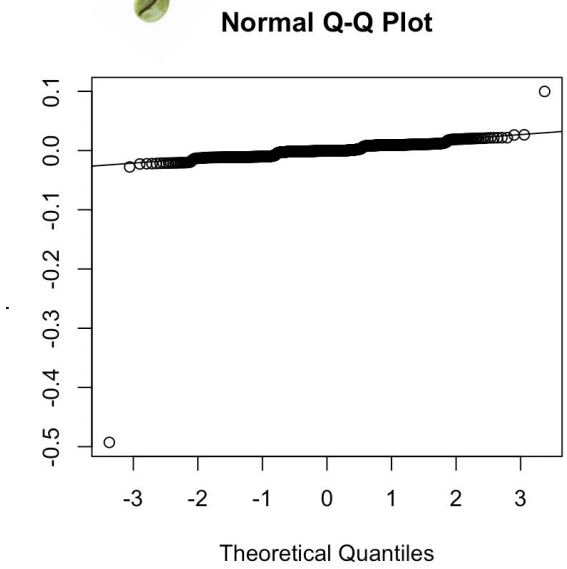


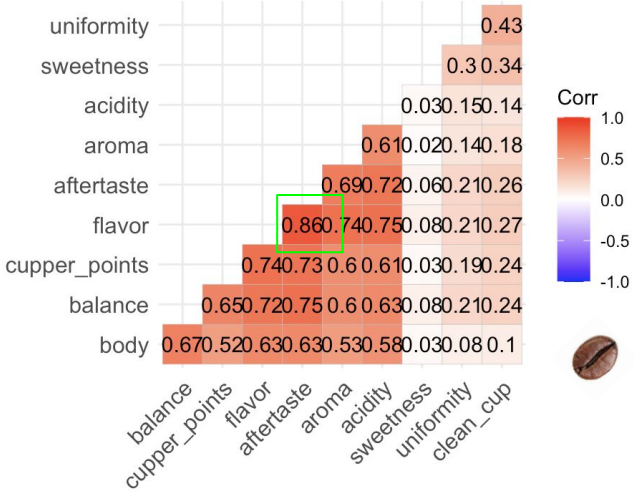
Variables	Mean	SD
Total Cupping Points	82.1512	2.669
Aroma	7.5724	0.3159
Flavor	7.5261	0.3414
Aftertaste	7.4066	0.3503
Acidity	7.5413	0.3192
Body	7.5231	0.3078
Balance	7.5236	0.3536
Uniformity	9.8422	0.4852
Cup Cleanliness	9.8423	0.7153
Sweetness	9.8640	0.5542
Cupper Points	7.508	0.4268

Assumptions

1. Linearity
2. Constant variance
3. Normality
4. Multicollinearity








```
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 ***
```

Modeling Results

^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 + 0.0161

- Flavor was the least reliable predictor
- Sweetness and cup cleanliness was the most reliable predictor.

```
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
```

Conclusion

- The p-value = < 2.2e-16.
- We reject the null hypothesis.
- All the coefficient β's are significant and have a correlation with the total cupping score.

References

LeDoux, James, Coffee ratings (2020), GitHub repository,

https://github.com/rfordatascience/tidytuesday/tree/master/data/2020/2020-07-07#coffee_ratingscsv

Thank you!