Why does coffee taste so good?

—— Classification of coffee beans quality

DAS Group 7

E Bolger, P L Lee, W Liao, Y Yang, Z Zhao

Introduction

- To investigate the relationship between the feature of coffees and quality.

- Using data sample (edited) with 892 observation and 9 variables including country, aroma, flavour, acidity, defects, altitude, harvested, quality-class and continent.

Methodology

STEP 1

- Having quick look at the datasets
- Removing data error N/A
- Added `continent` variable
- Converting to tidy data format

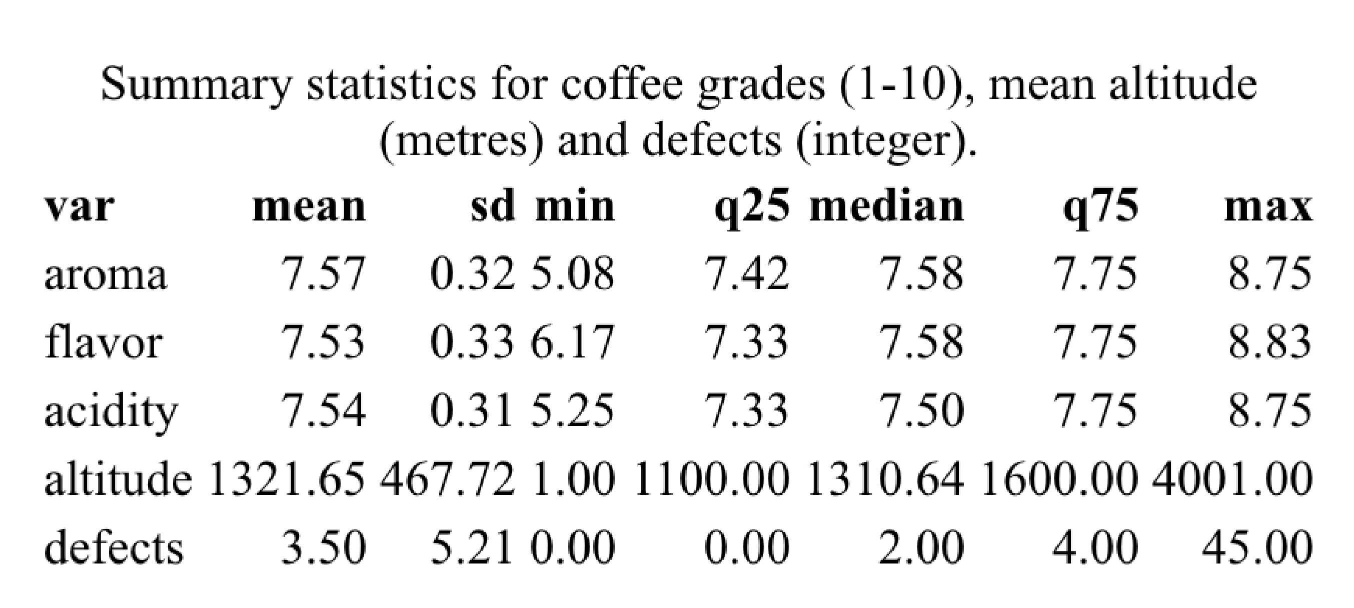
STEP 2

- Fitting generalized linear model with 3 link function
- Model Selection
- Verify Assumptions

STEP 3

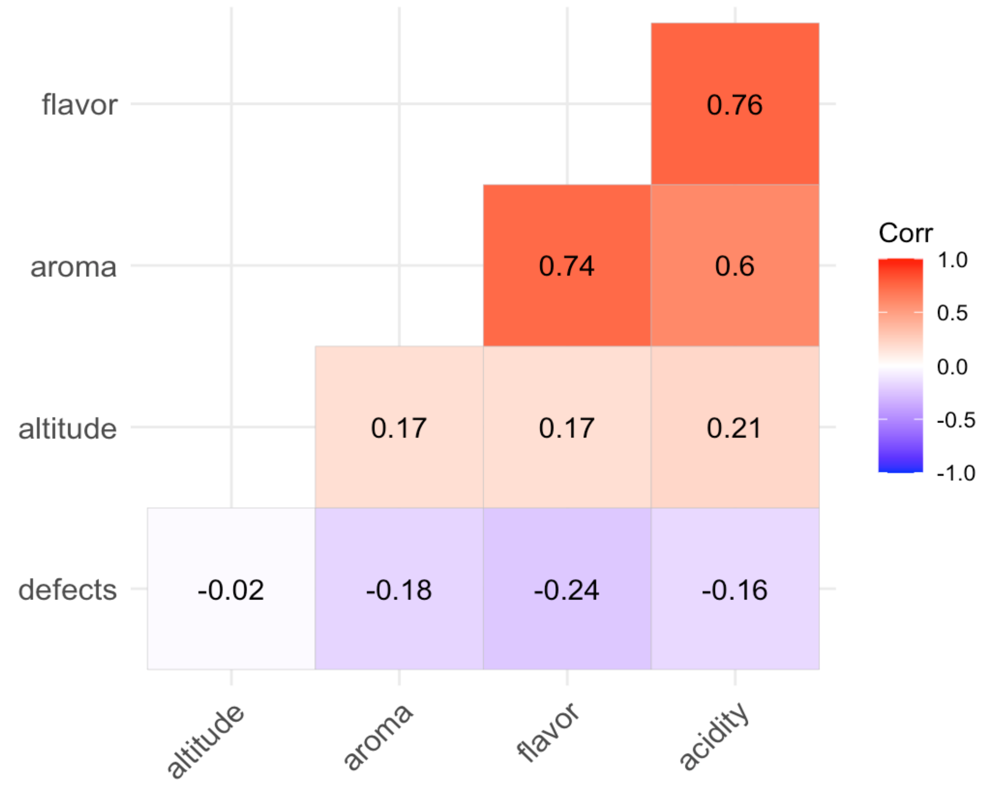
- Draw conclusions
- Further research and considerations

Exploratory Data Analysis (1)

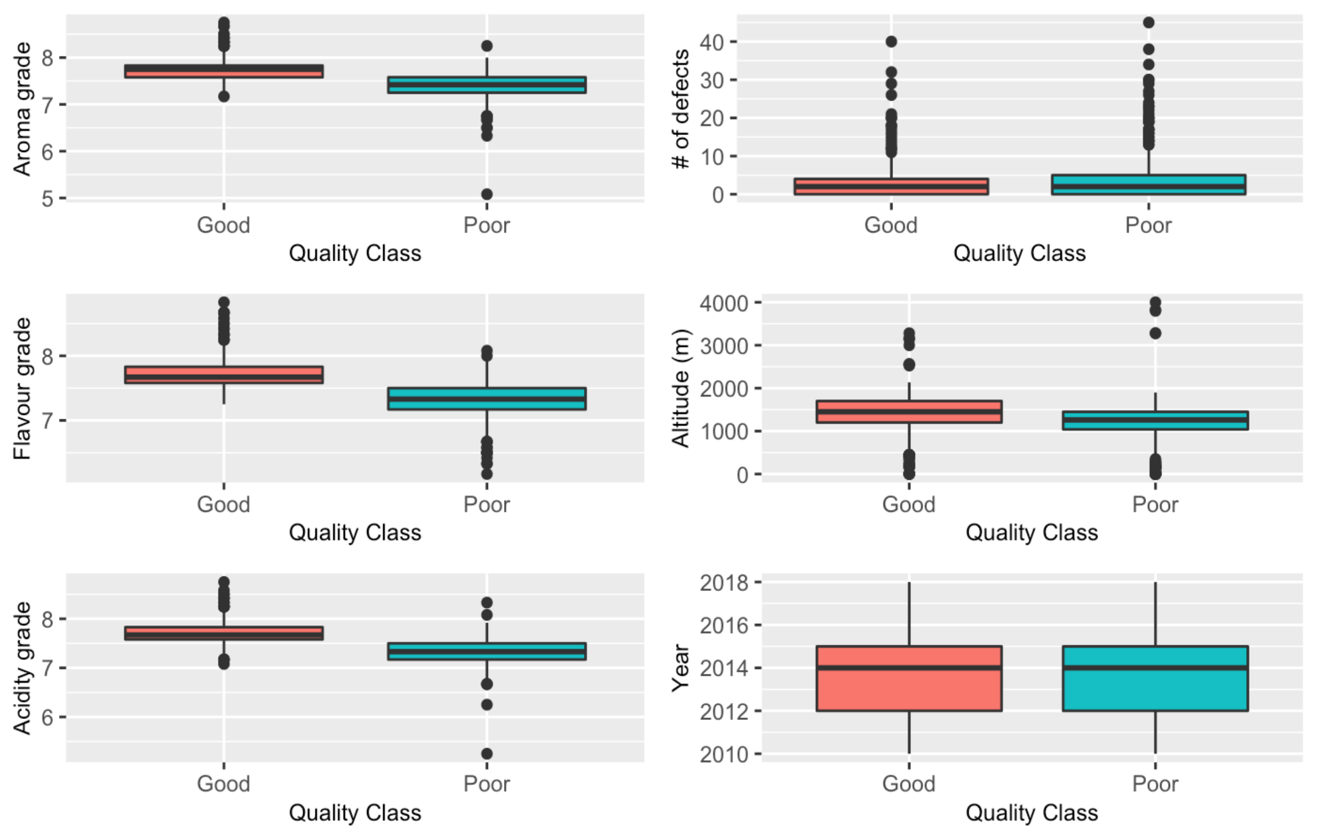


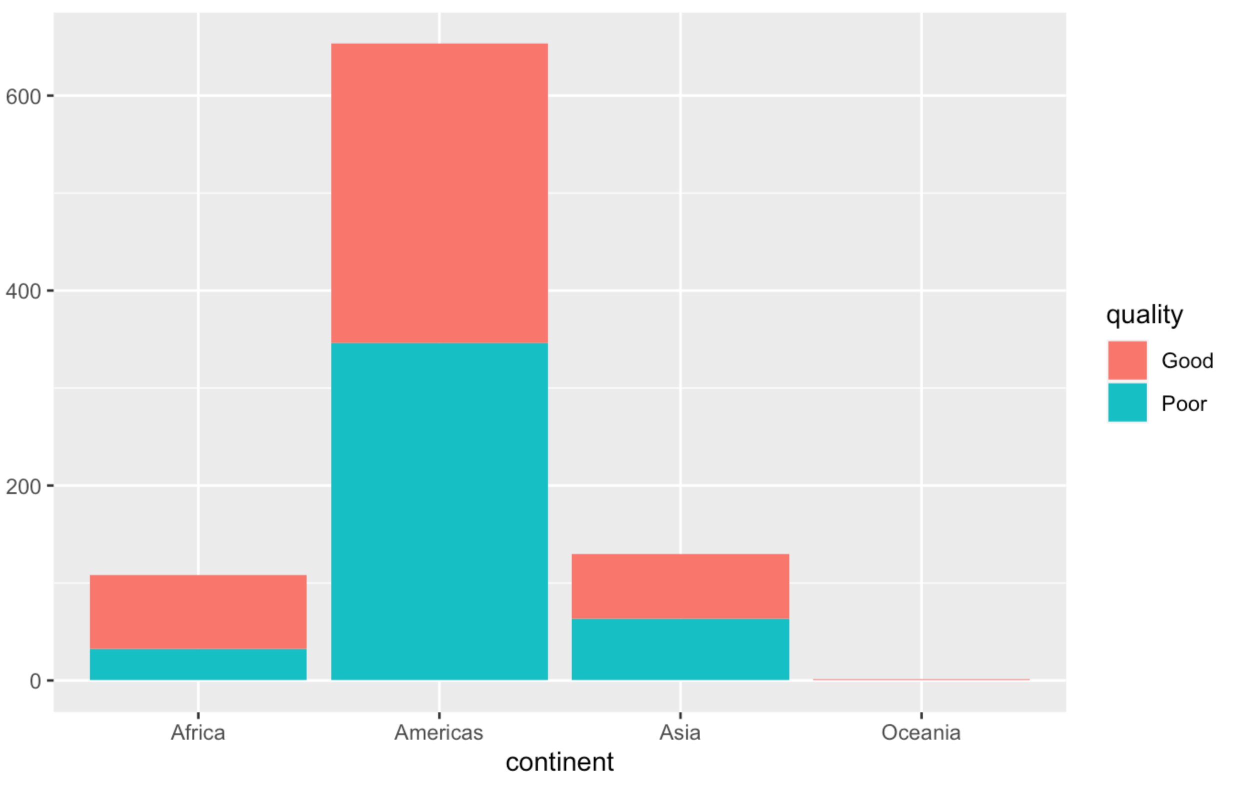
Summary statistics for coffee grades (1-10), mean altitude (metres) and defects (integer).

var	mean	sd	min	q25	median	q75	max
aroma	7.57	0.32	5.08	7.42	7.58	7.75	8.75
flavor	7.53	0.33	6.17	7.33	7.58	7.75	8.83
acidity	7.54	0.31	5.25	7.33	7.50	7.75	8.75
altitude	1321.65	467.72	1.00	1100.00	1310.64	1600.00	4001.00
defects	3.50	5.21	0.00	0.00	2.00	4.00	45.00

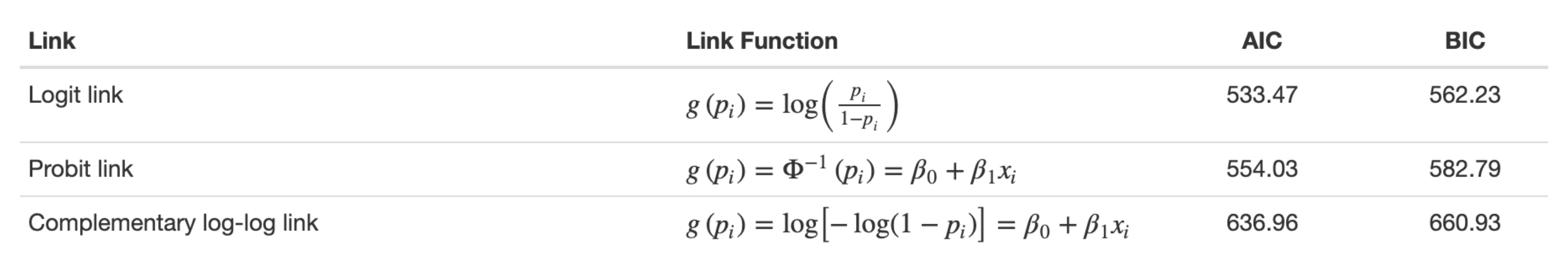


Exploratory Data Analysis (2)



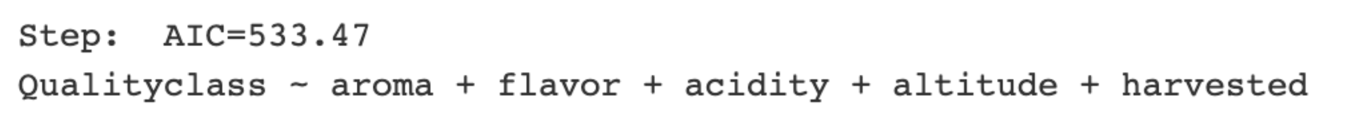


Data Analysis



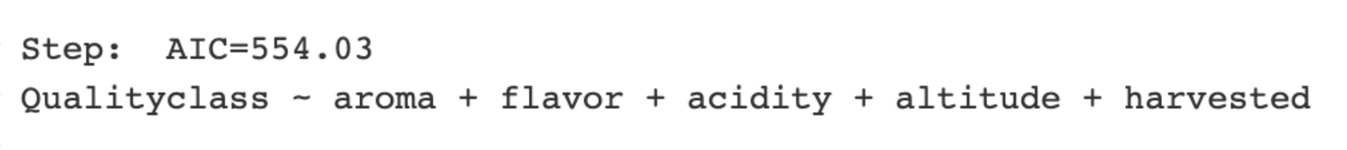
Link	Link Function	AIC	BIC
Logit link	$g\left(p_i\right) = \log\left(\frac{p_i}{1-p_i}\right)$	533.47	562.23
Probit link	$g\left(p_i\right) = \Phi^{-1}\left(p_i\right) = \beta_0 + \beta_1 x_i$	554.03	582.79
Complementary log-log link	$g\left(p_i\right) = \log\left[-\log(1-p_i)\right] = \beta_0 + \beta_1 x_i$	636.96	660.93

Logit Link



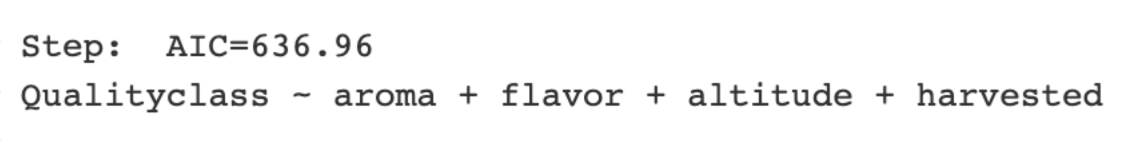
```
Step:  AIC=533.47
Qualityclass ~ aroma + flavor + acidity + altitude + harvested
```

Probit Link



```
Step:  AIC=554.03
Qualityclass ~ aroma + flavor + acidity + altitude + harvested
```

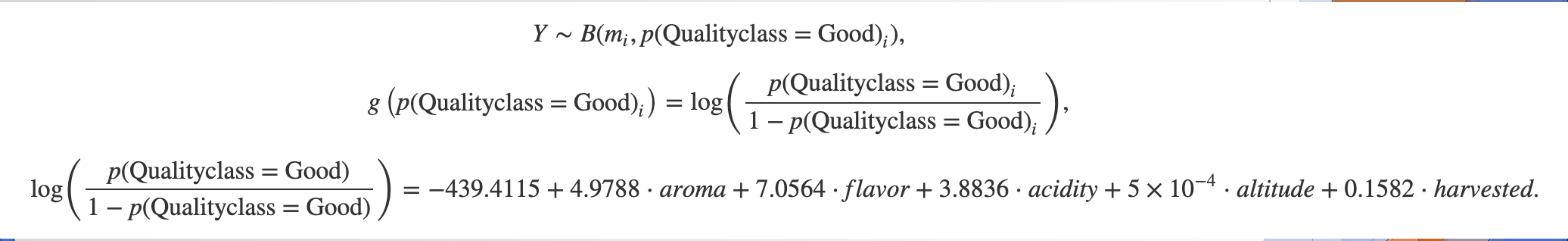
Complementary log-log Link



```
Step:  AIC=636.96
Qualityclass ~ aroma + flavor + altitude + harvested
```

Model Selection

The logit link function was chosen because it had the lowest AIC & BIC



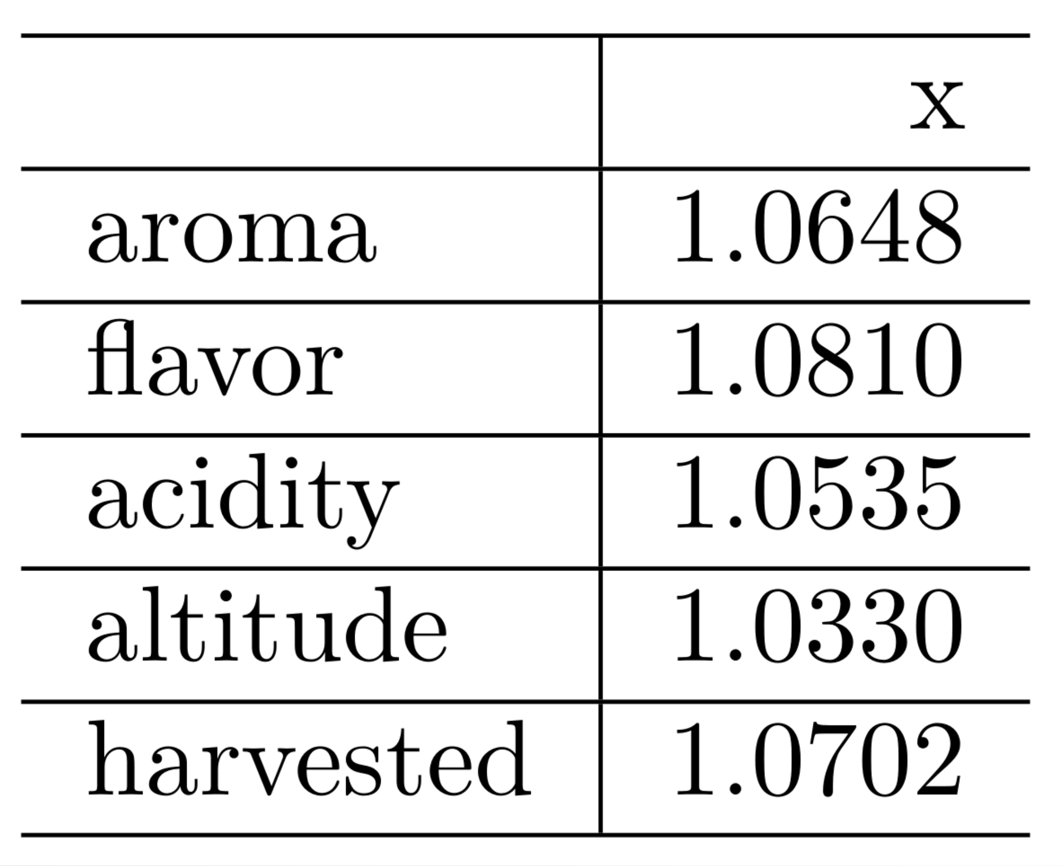
$$Y \sim B(m_i, p(\text{Qualityclass} = \text{Good})_i),$$

$$g\left(p(\text{Qualityclass} = \text{Good})_i\right) = \log\left(\frac{p(\text{Qualityclass} = \text{Good})_i}{1 - p(\text{Qualityclass} = \text{Good})_i}\right),$$

$$\log\left(\frac{p(\text{Qualityclass} = \text{Good})}{1 - p(\text{Qualityclass} = \text{Good})}\right) = -439.4115 + 4.9788 \cdot \mathit{aroma} + 7.0564 \cdot \mathit{flavor} + 3.8836 \cdot \mathit{acidity} + 5 \times 10^{-4} \cdot \mathit{altitude} + 0.1582 \cdot \mathit{harvested}.$$

Model Assumptions

The result shows that all VIF of the variables are small.



	x
aroma	1.0648
flavor	1.0810
acidity	1.0535
altitude	1.0330
harvested	1.0702

Conclusion and Further Work

- We can **conclude** that all variable have positive correlation with the coffee quality.
- Address this problem with other methods such as tree model. Divide the data into training set and test set. Measure coffee quality with quantitative data rather than binary variable.
- Explore more variable that might affect the coffee quality.