Group 18

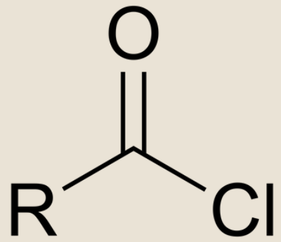
Red Wine Quality

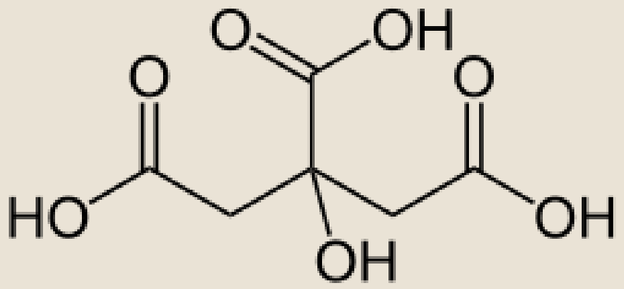
Inferential Analysis

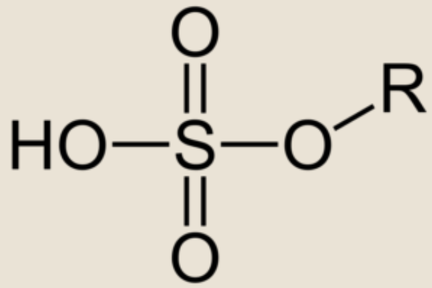
Bombini Nicola
Cataldo Giulia
Iania Leonardo
Signò Jacopo

Goal of our analysis

Investigating how physiochemical properties affect the quality of the Portuguese "Vinho Verde" red wine.

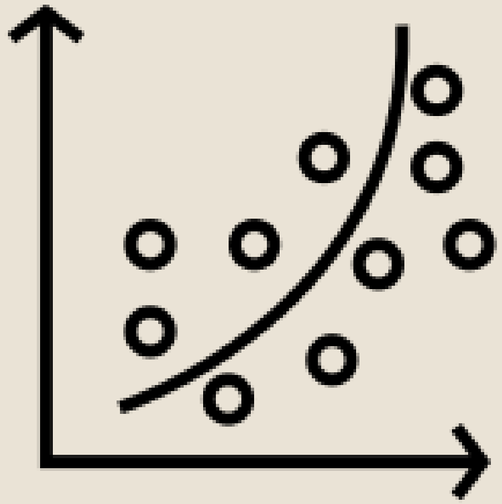






Models:

- Ordered multinomial model
- Binomial model



Conclusions:

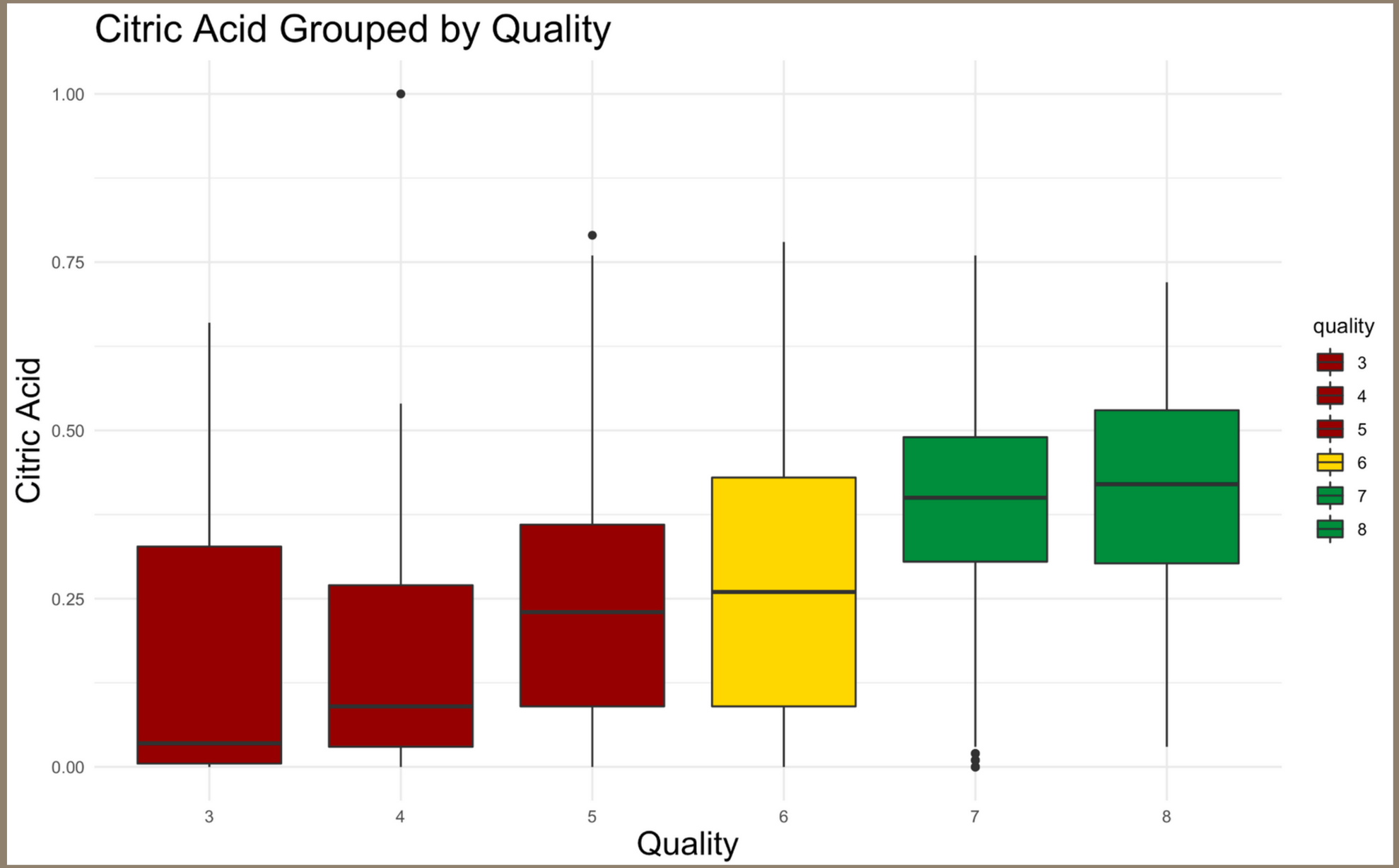
The two models lead us to very similar results

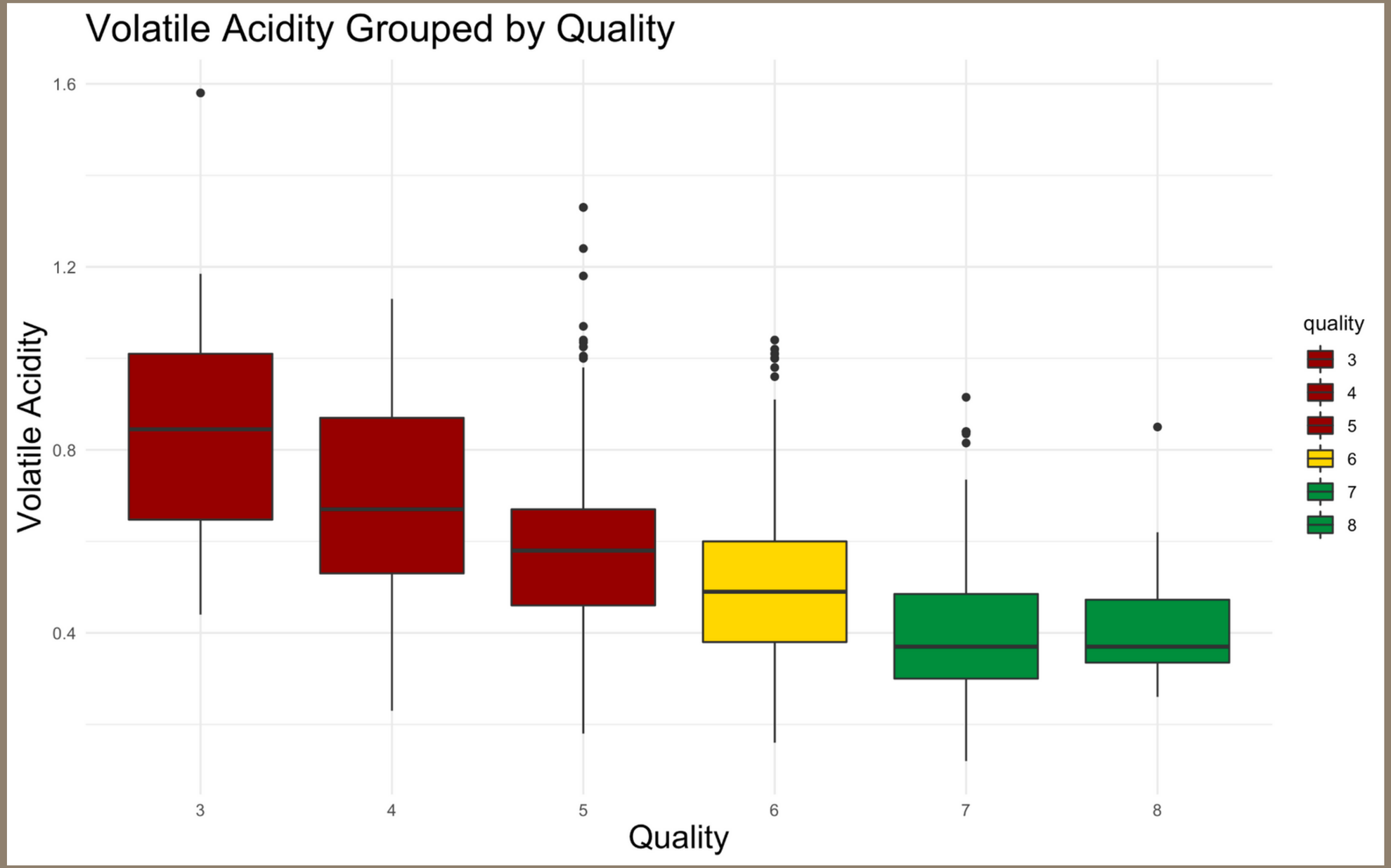
Dataset

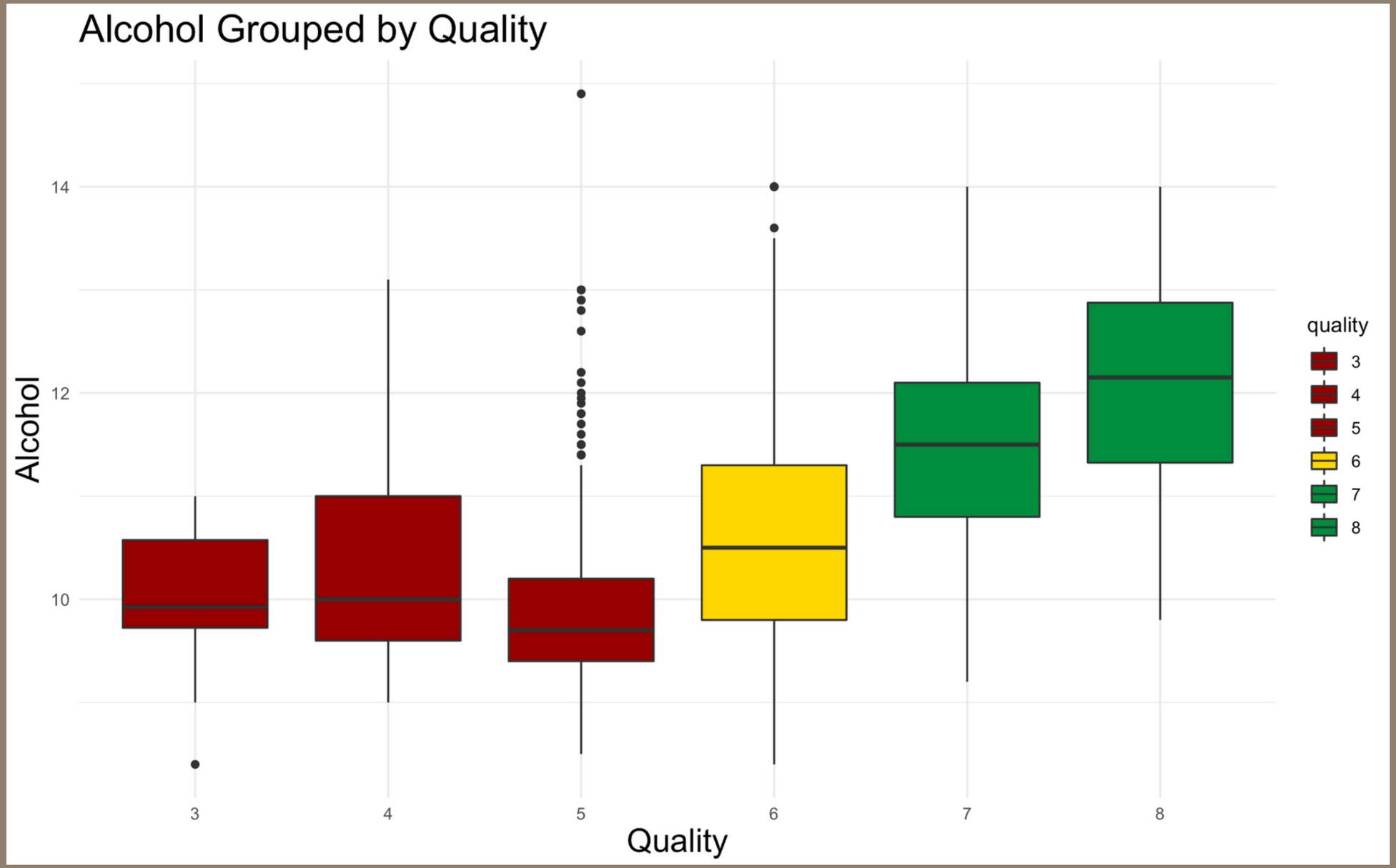
11 numerical variables
1 categorical response

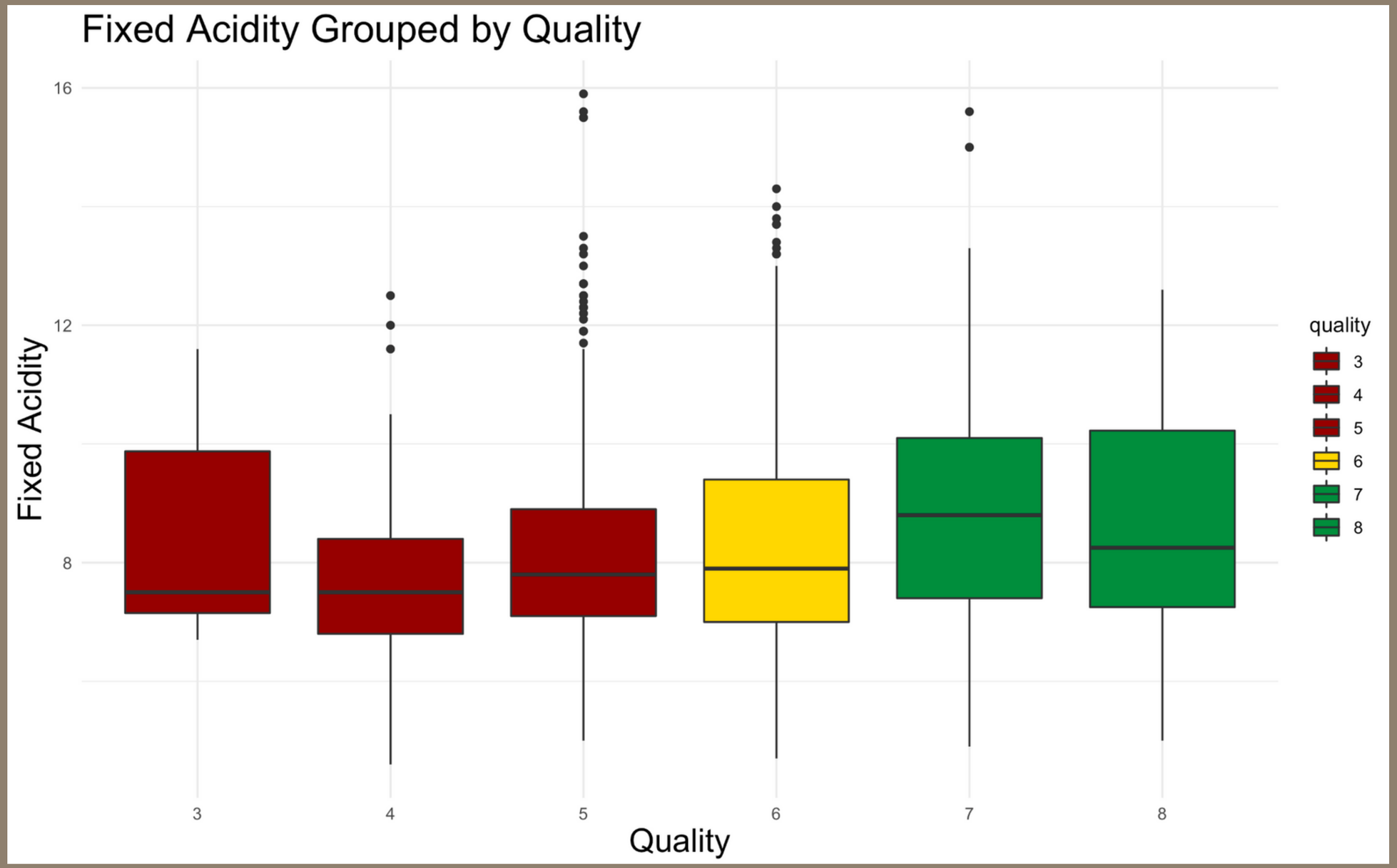
- **Quality**: quality on a sale from 3 to 8
- **Fixed acidity**: most of the acids in wine
- **Volatile acidity**: the amount of acetic acid, which in high quantities can lead to an unpleasant vinegar taste
- **Citric acid**: found in small quantities, it adds freshness and flavour to wine
- **Residual sugar**: sugars remaining after fermentation stops

- **Chlorides**: the amount of salt in wine
- **Free sulfur dioxide**: prevents microbial growth and fermentation of wine
- **Total sulfur dioxide**: sum of free and bound sulfur dioxide, in high concentrations it becomes evident in taste and nose
- **Density**: close to density of water
- **pH**: how acid the wine is. Most wines are between 3 and 4
- **Sulphates**: added to wine because it's an antimicrobilal and antioxidant
- **Alcohol**: amount of alcohol in the wine

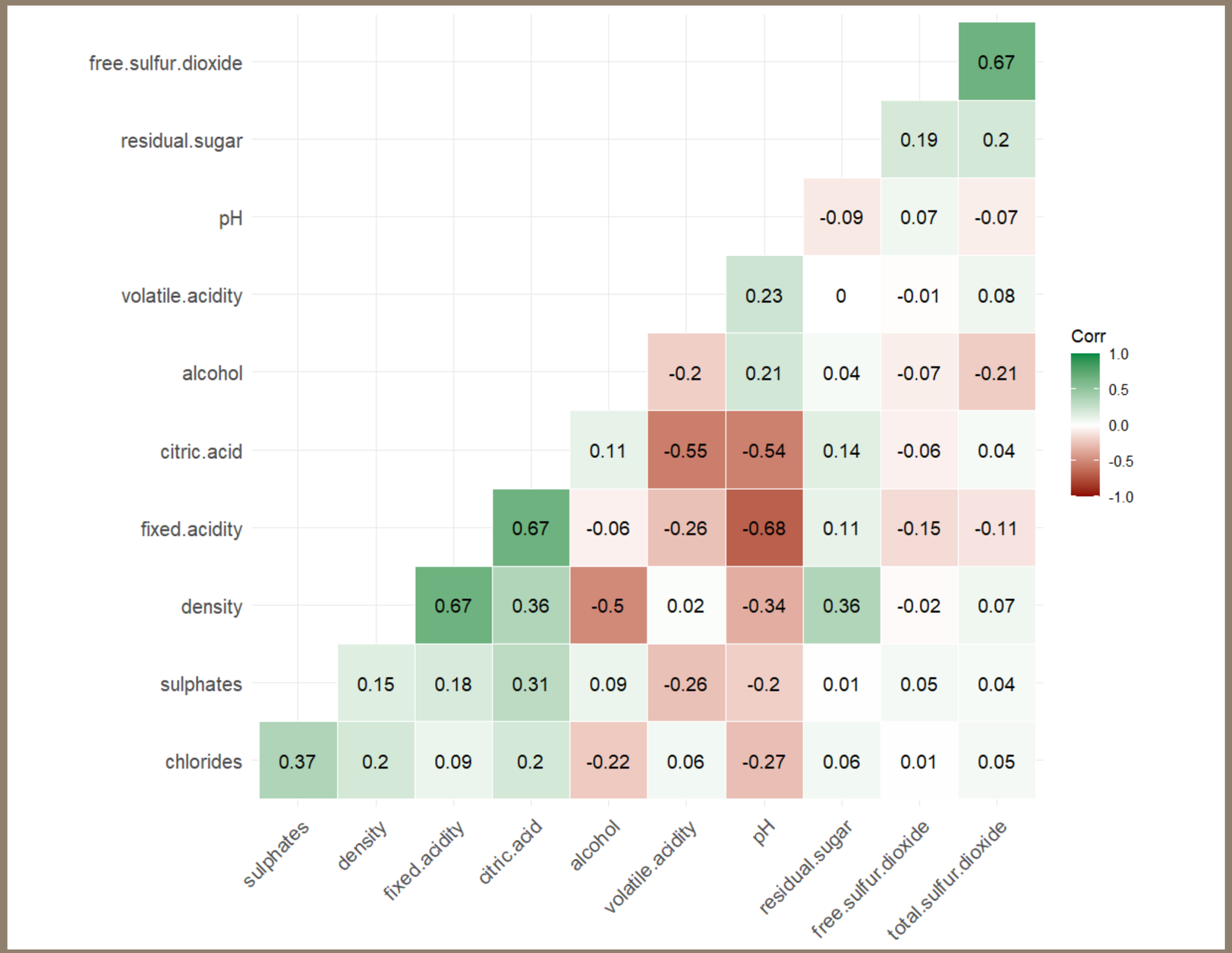

Citric Acid Grouped by Quality
Citric Acid
1.00
0.75
0.50
0.25
0.00
3
4
5
6
7
8
Quality
quality
3
4
5
6
7
8


Volatile Acidity Grouped by Quality
Volatile Acidity
1.6
1.2
0.8
0.4
3
4
5
6
7
8
Quality
quality
3
4
5
6
7
8


Alcohol Grouped by Quality
Alcohol
14
12
10
3
4
5
6
7
8
Quality
quality
3
4
5
6
7
8

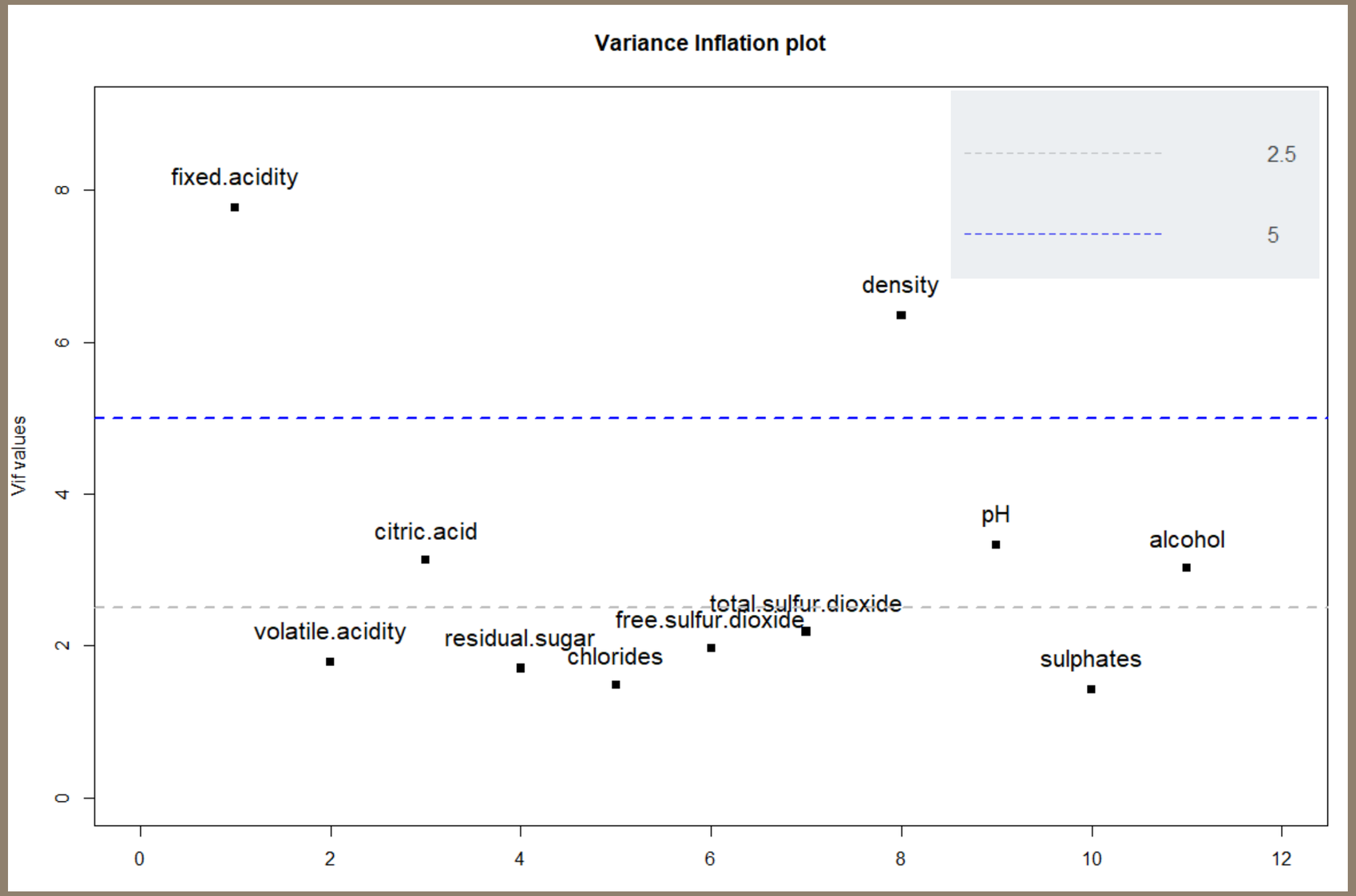

Fixed Acidity Grouped by Quality
16
12
8
Fixed Acidity
3
4
5
6
7
8
Quality
quality
3
4
5
6
7
8

Correlation matrix with initial variables



VIF

Fixed acidity and density above the VIF=5 conventional threshold



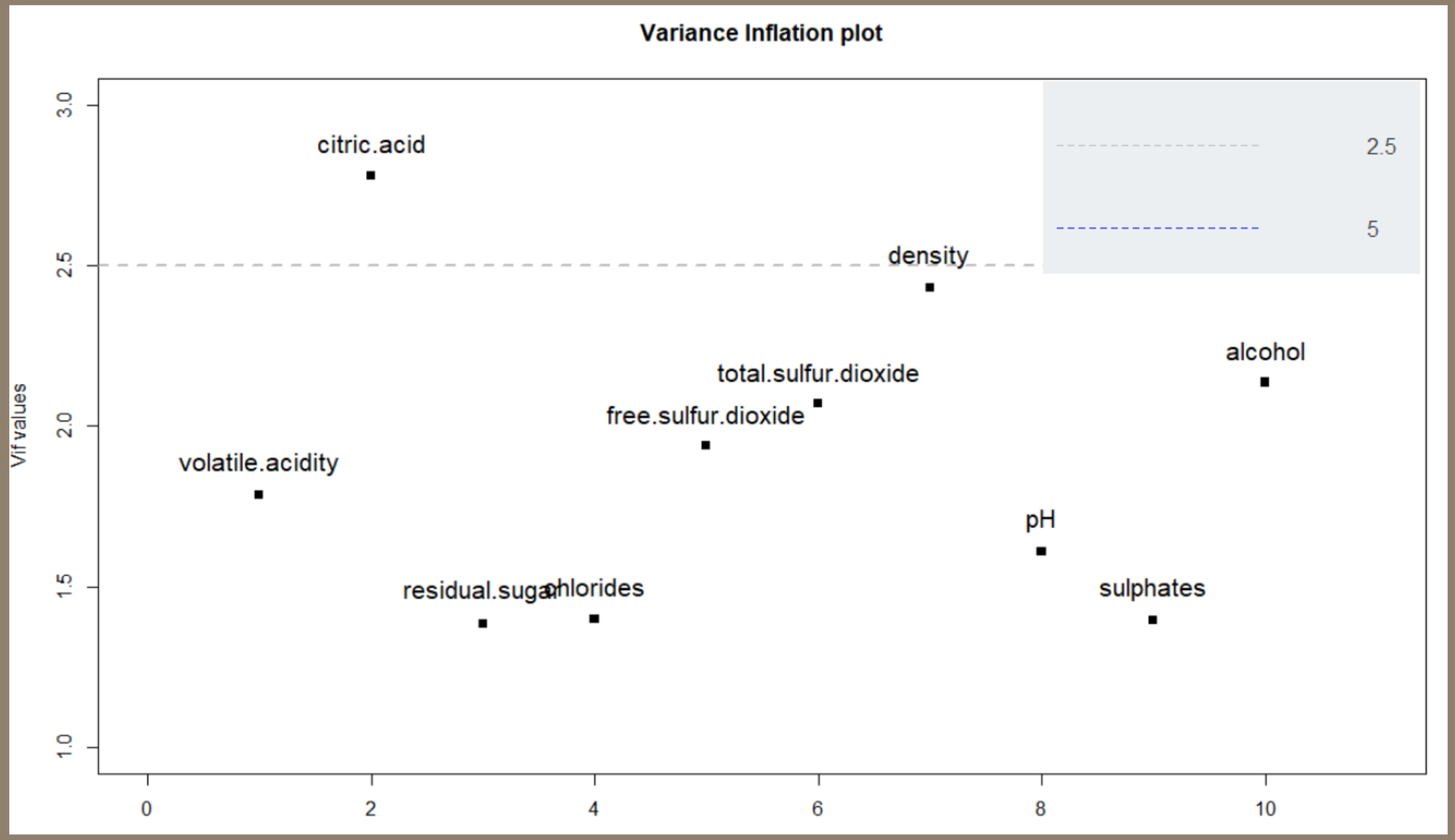
Variable selection

Fixed Acidity vs All other variables: R-squared 0.8713

Total sulfur dioxide - free sulfur dioxide ⟶ Bound sulfur dioxide

VIF

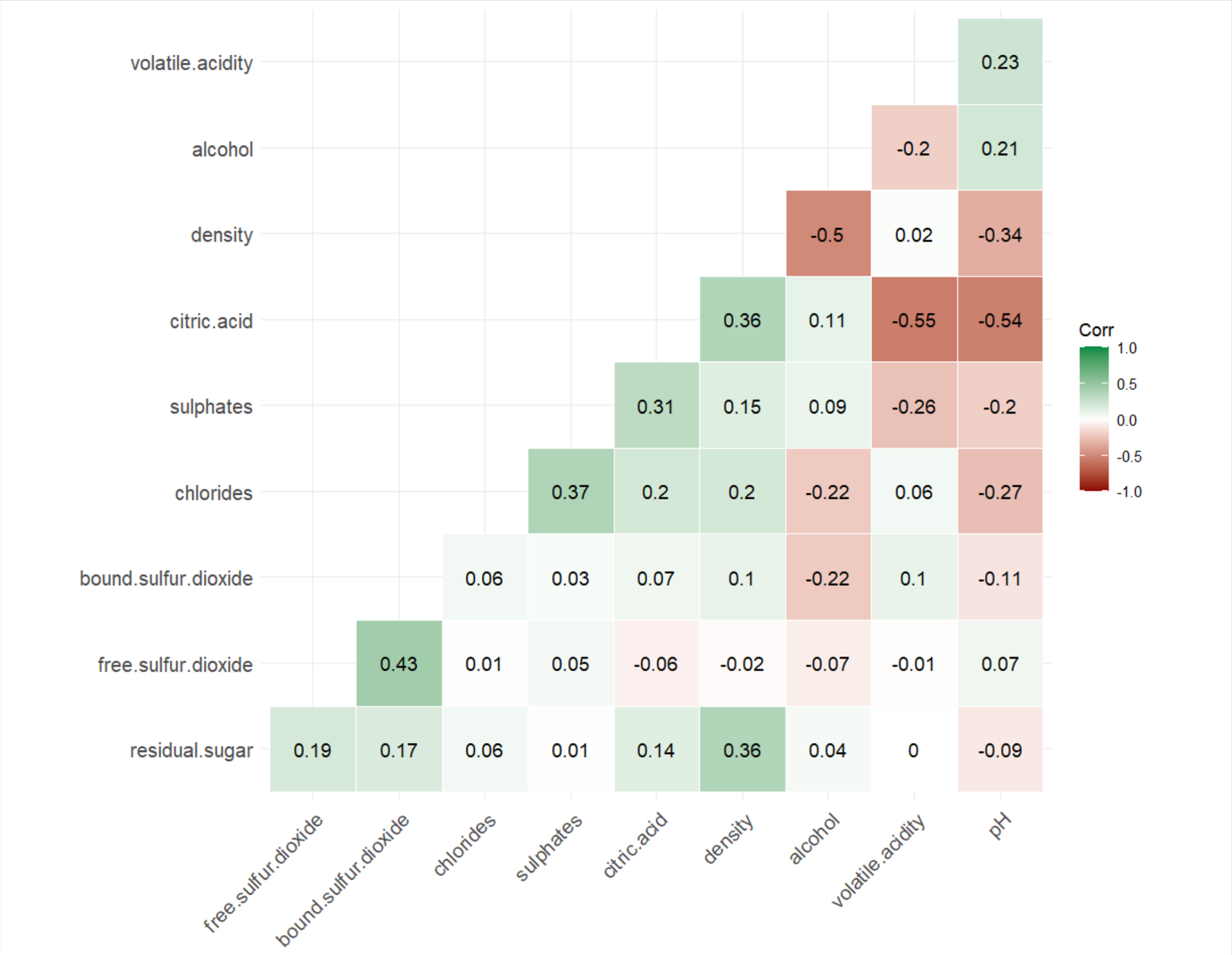
No problems anymore with any variable



Variable selection

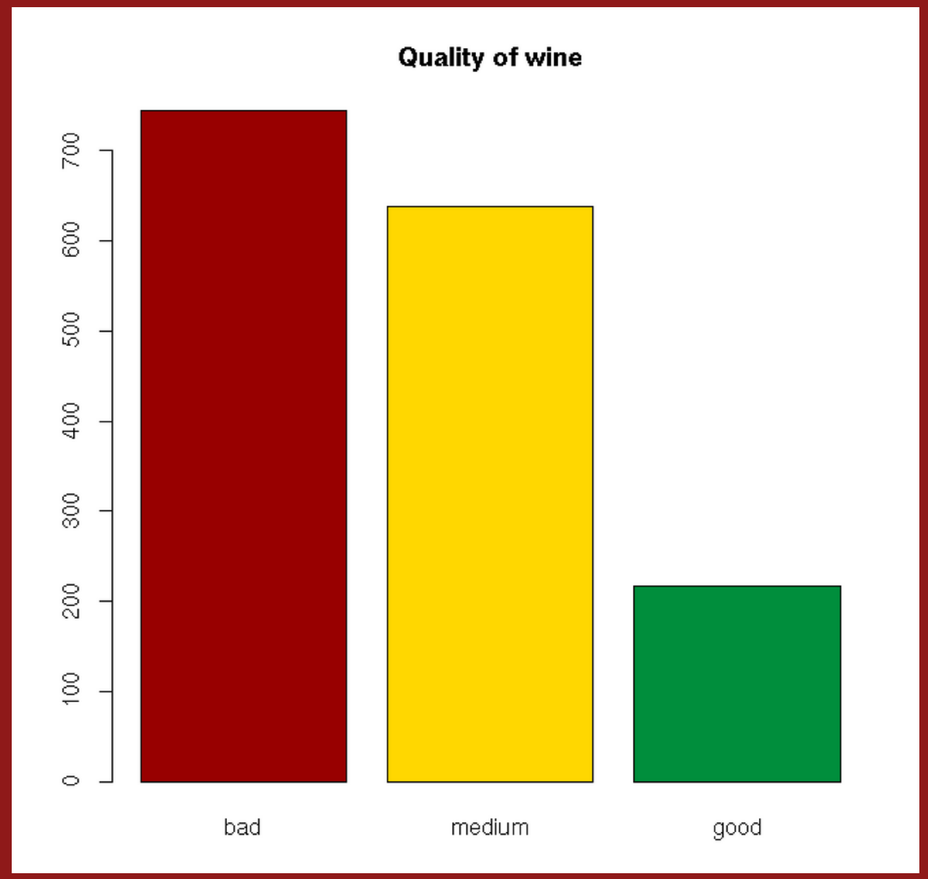
Fixed Acidity vs All other variables: R-squared 0.8713

Total sulfur dioxide - free sulfur dioxide ⟶ Bound sulfur dioxide



Correlation matrix with initial variables

Ordered Multinomial Model



Grades 3, 4 and 5 → bad

Grade 6 → medium

Grades 7 and 8 → good

Mathematical aspects

$$\gamma_{ij} = P(Y_i \leq j)$$

Cumulative probabilities

$$g(\gamma_{ij}) = \theta_j - \beta^T x_i$$

General form of the link function

$$\log\left(\frac{\gamma_{ij}}{1-\gamma_{ij}}\right)$$

Log-odds function

Final model

$$logit(P(\text{bad})) = 5.17 - 2.80x_1 - 5.91x_2 - 1.04x_3 + 3.01x_4 + 0.90x_5 - 0.02x_6$$

$$logit(P(\text{medium or bad})) = 8.06 - 2.80x_1 - 5.91x_2 - 1.04x_3 + 3.01x_4 + 0.90x_5 - 0.02x_6$$

Coefficients:

		Value	Std. Error	t value	p value
x_1	volatile.acidity	-2.80816	0.34939	-8.037	9.189600e-16
x_2	chlorides	-5.91108	1.37260	-4.306	1.658703e-05
x_3	pH	-1.04281	0.38016	-2.743	6.087140e-03
x_4	sulphates	3.01230	0.36389	8.278	1.252269e-16
x_5	alcohol	0.90718	0.05916	15.333	4.584314e-53
x_6	bound.sulfur.dioxide	-0.01761	0.00241	-7.308	2.716690e-13

Intercepts:

		Value	Std. Error	t value	p value
θ_1	bad|medium	5.1728	1.3022	3.9724	7.113811e-05
θ_2	medium|good	8.0583	1.3144	6.1309	8.736129e-10

Residual Deviance: 2457.885
AIC: 2473.885

Model Selection

Criterion-based test:

- AIC
- BIC

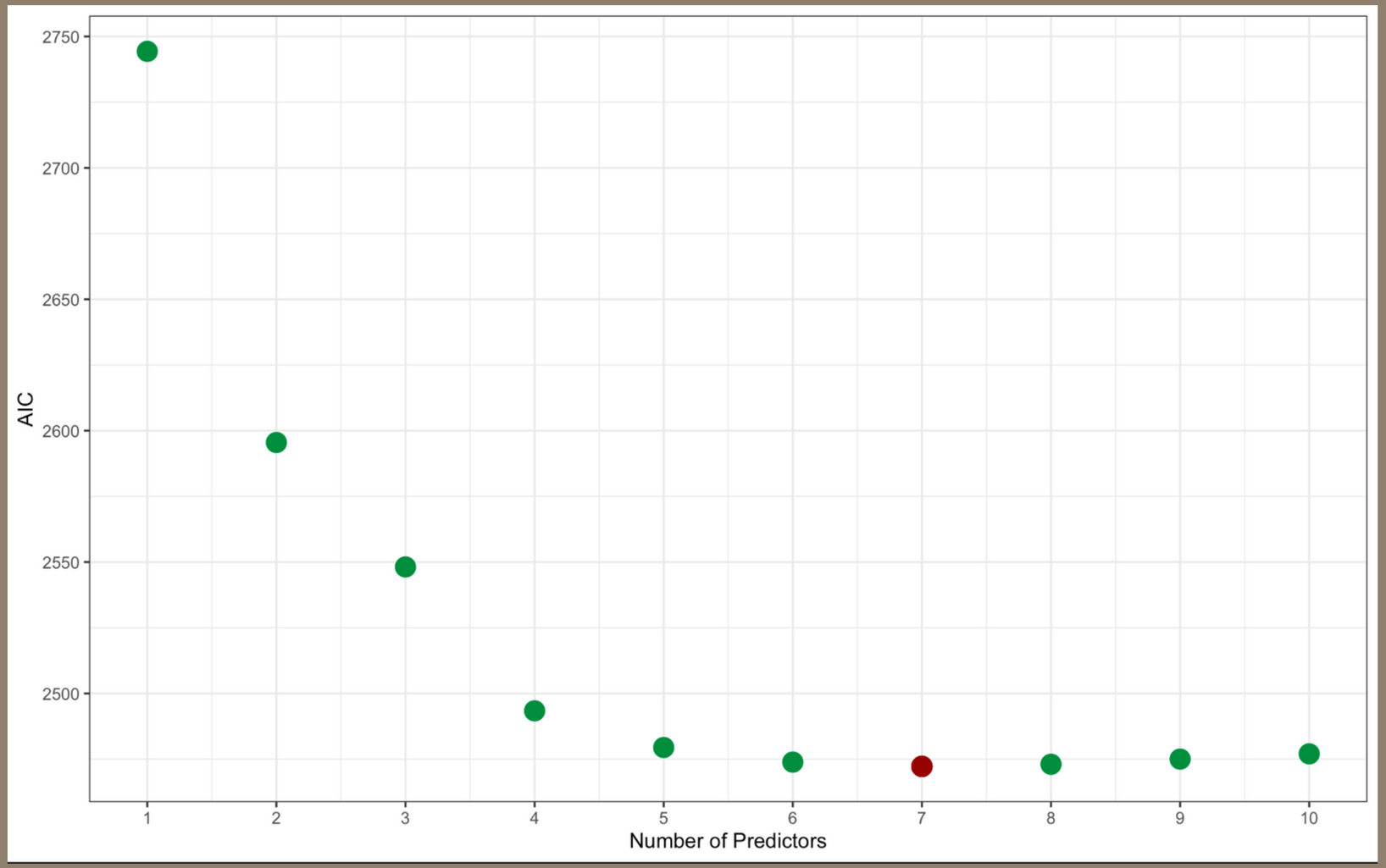
Testing-based procedures:

- Single term deletion with "Chi Square" test

Comparison with the full model:

ANOVA with X^2 test

AIC



7 predictors

- volatile acidity
- chlorides
- pH
- sulphates
- alcohol
- bound sulphur dioxide
- residual sugar

Alternative model

with 7 variables

Coefficients:

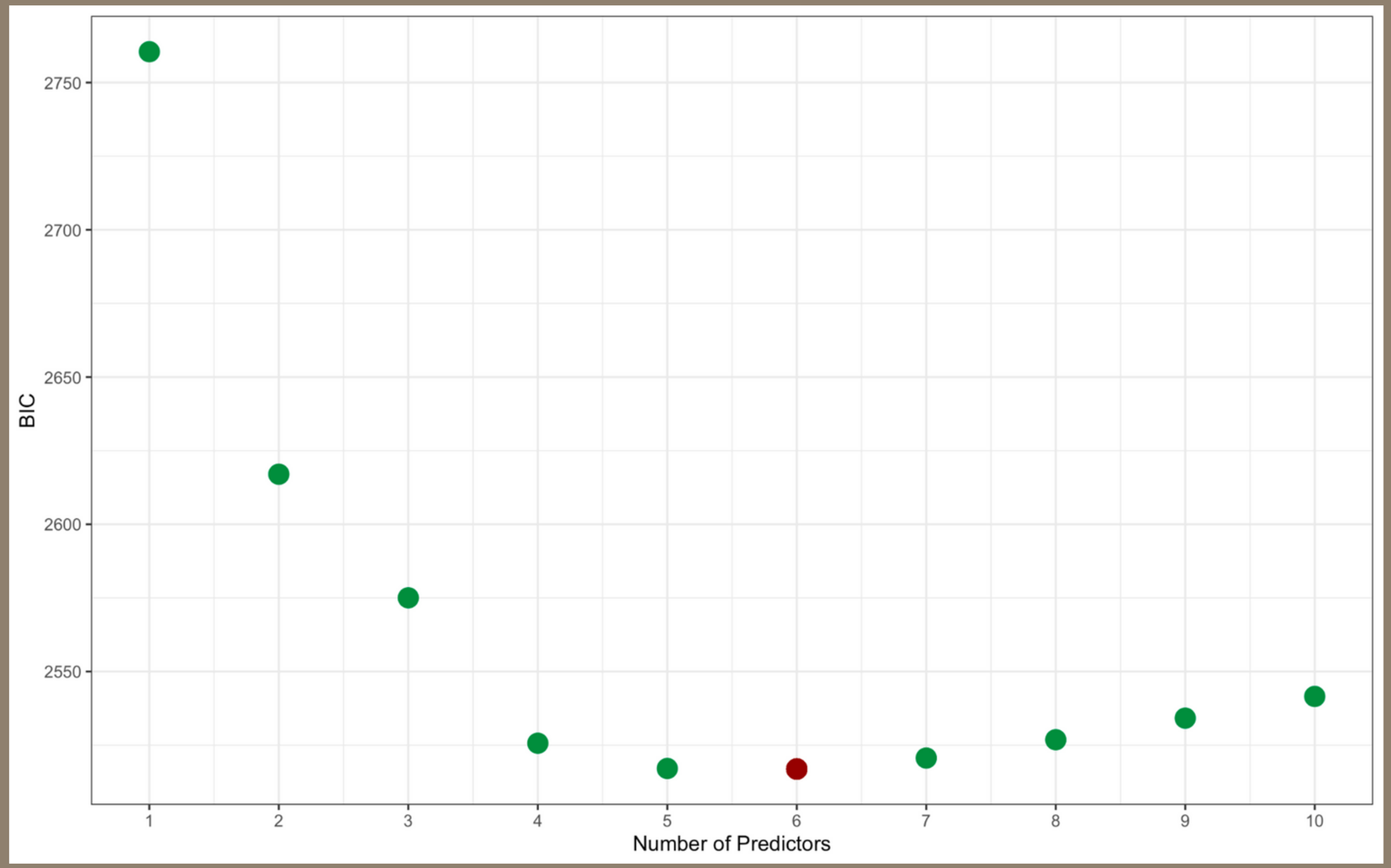
		Value	Std. Error	t value	p value
x_1	volatile.acidity	-2.81720	0.350156	-8.046	0.000
x_2	chlorides	-6.09406	1.387370	-4.393	0.000
x_3	pH	-0.99637	0.381419	-2.612	0.009
x_4	sulphates	3.05175	0.364630	8.369	0.000
x_5	alcohol	0.89775	0.059351	15.126	0.000
x_6	bound.sulfur.dioxide	-0.01851	0.002471	-7.492	0.000
x_7	residual.sugar	0.07500	0.038838	1.931	0.053

Intercepts:

		Value	Std. Error	t value	p value
θ_1	bad|medium	5.3962	1.3093	4.1215	0.000
θ_2	medium|good	8.2849	1.3216	6.2687	0.000

Residual Deviance: 2454.24
AIC: 2472.24

BIC



6 predictors

- volatile acidity
- chlorides
- pH
- sulphates
- alcohol
- bound sulphur dioxide

Stepwise elimination

Based on "Chi square" test

```
Single term deletions

Model:
quality ~ volatile.acidity + chlorides + pH + sulphates + alcohol +
    bound.sulfur.dioxide + residual.sugar
                      Df    AIC     LRT  Pr(>Chi)
<none>                     2472.2
volatile.acidity       1 2539.0  68.757 < 2.2e-16 ***
chlorides              1 2491.3  21.084 4.396e-06 ***
pH                     1 2477.1   6.846  0.008884 **
sulphates              1 2544.4  74.127 < 2.2e-16 ***
alcohol                1 2727.9 257.660 < 2.2e-16 ***
bound.sulfur.dioxide   1 2534.0  63.797 1.379e-15 ***
residual.sugar         1 2473.9   3.645  0.056229 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The test suggests using a model with 6 variables

Statistical tests

"Chi square" test of LR

	Resid. df	Resid. Dev	Test	Df	LR stat.	Pr(Chi)
Final model	1591	2457.885				
Total model	1587	2453.016	1 vs 2	4	4.869737	0.3009263

H0: the parameters we removed are all equal to 0

We can remove these variables:

- Citric acid
- Free sulphur dioxide
- Density
- Residual sugar

Model diagnostic

Proportional odds assumption:

Test for	X2	df	probability
Omnibus	7.36	6	0.29
volatile.acidity	0.51	1	0.47
chlorides	1.36	1	0.24
pH	2.88	1	0.09
sulphates	1.08	1	0.3
alcohol	1.29	1	0.26
bound.sulfur.dioxide	0.05	1	0.82

H0: Parallel Regression Assumption holds

Brant test

$$\begin{cases} \mathrm{H}_0 : \ \beta_k = \beta \\ \mathrm{H}_1 : \ \beta_k = \phi_k \beta \end{cases}$$

$$logit\left(\gamma_1(x)\right) - logit\left(\gamma_2(x)\right) = \theta_1 - \theta_2$$

Model diagnostic

Test for goodness of fit: Hosmer-Lemeslow test

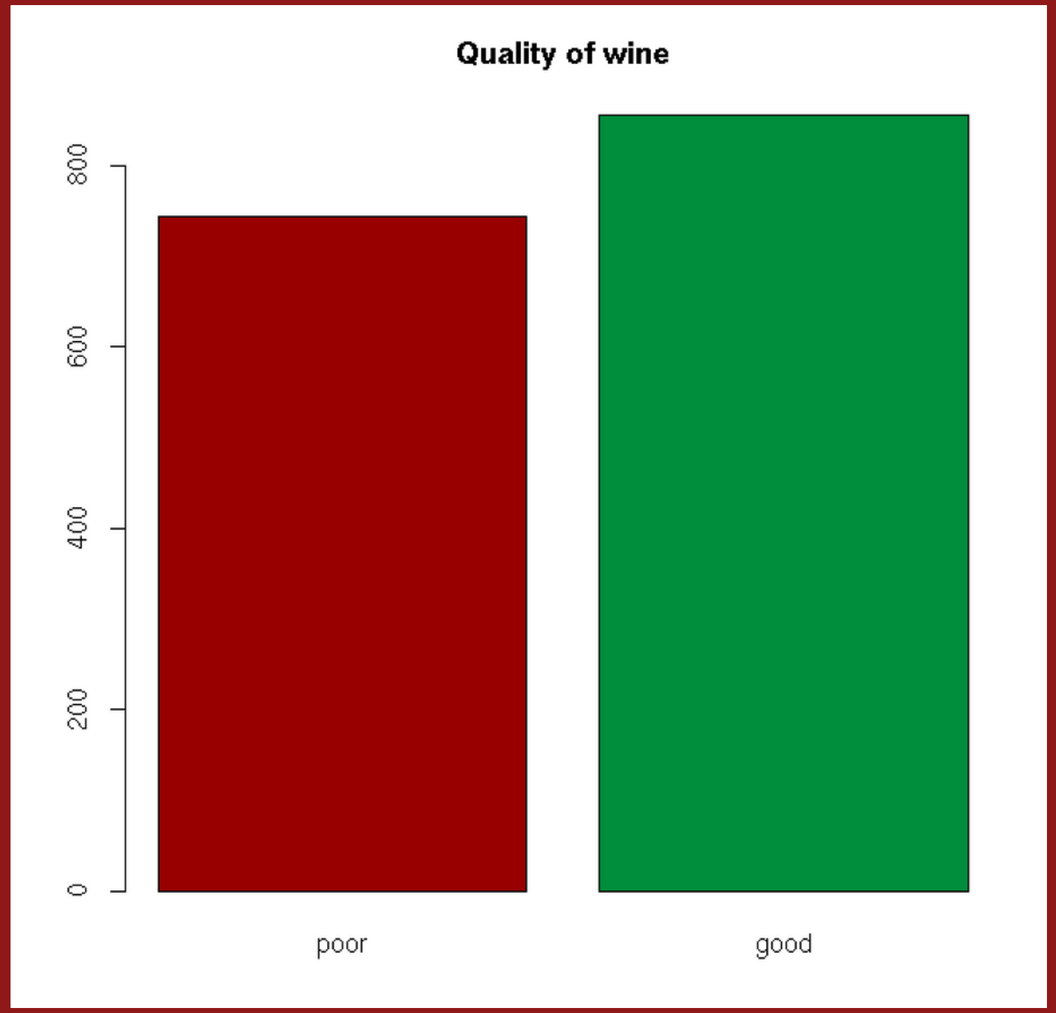
```
        Hosmer and Lemeshow test (ordinal model)

data:  data2$quality, fitted(poltt)
X-squared = 24.835, df = 17, p-value = 0.09847
```

$$\begin{cases} \boldsymbol{H_0: Model\ suits\ the\ data} \\ \boldsymbol{H_1: Model\ doesn't\ suit\ the\ data} \end{cases}$$

$$X^2_{LS} = \sum_{j=1}^{J} \frac{(y_j - m_j\hat{p}_j)^2}{m_j\hat{p}_j(1-\hat{p}_j)}$$

Binomial Model



Grades 3, 4, and 5 → bad
Grades 6, 7, and 8 → good

With the division we chose, both categories are well represented

Mathematical aspects

$$Y_i \sim Bin(n_i, p_i)$$

Binomial Distribution

$$P(Y_i = y_i) = \binom{n_i}{y_i} p_i^{y_i}(1 - p_i)^{n_i - p_i}$$

$$\eta = log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + .. + \beta_q x_q$$

Logit Link Model

$$p = g^{-1}(\eta) = \frac{e^\eta}{e^\eta + 1}$$

Inverse of Logit

Final Model

$$\eta_i = -8.10 - 2.91x_1 - 4.45x_2 + 0.86x_3 + 2.71x_4 - 0.02x_5$$

```
   Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
   (Intercept)            -8.09506    0.80702 -10.031  < 2e-16 ***
x1 volatile.acidity       -2.91314    0.37034  -7.866 3.66e-15 ***
x2 chlorides              -4.45501    1.43196  -3.111  0.00186 **
x3 alcohol                 0.86127    0.07078  12.168  < 2e-16 ***
x4 sulphates               2.71473    0.42875   6.332 2.43e-10 ***
x5 bound.sulfur.dioxide   -0.01639    0.00244  -6.718 1.84e-11 ***
   ---

       Null deviance: 2209.0  on 1598  degrees of freedom
   Residual deviance: 1665.7  on 1593  degrees of freedom
```

Model Selection

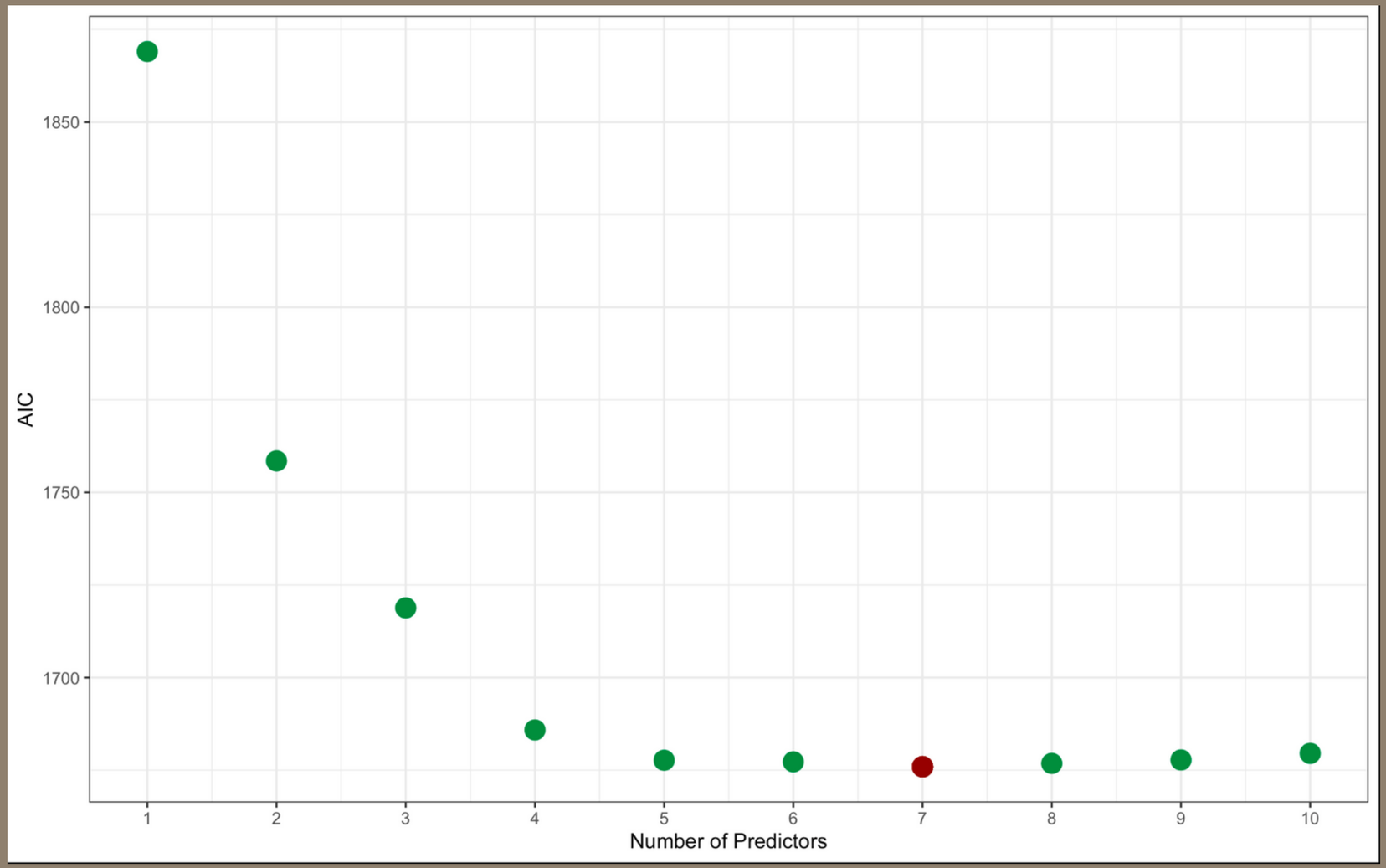
Criterion-based test:

- AIC
- BIC

Testing-based procedures:

- Backward elimination

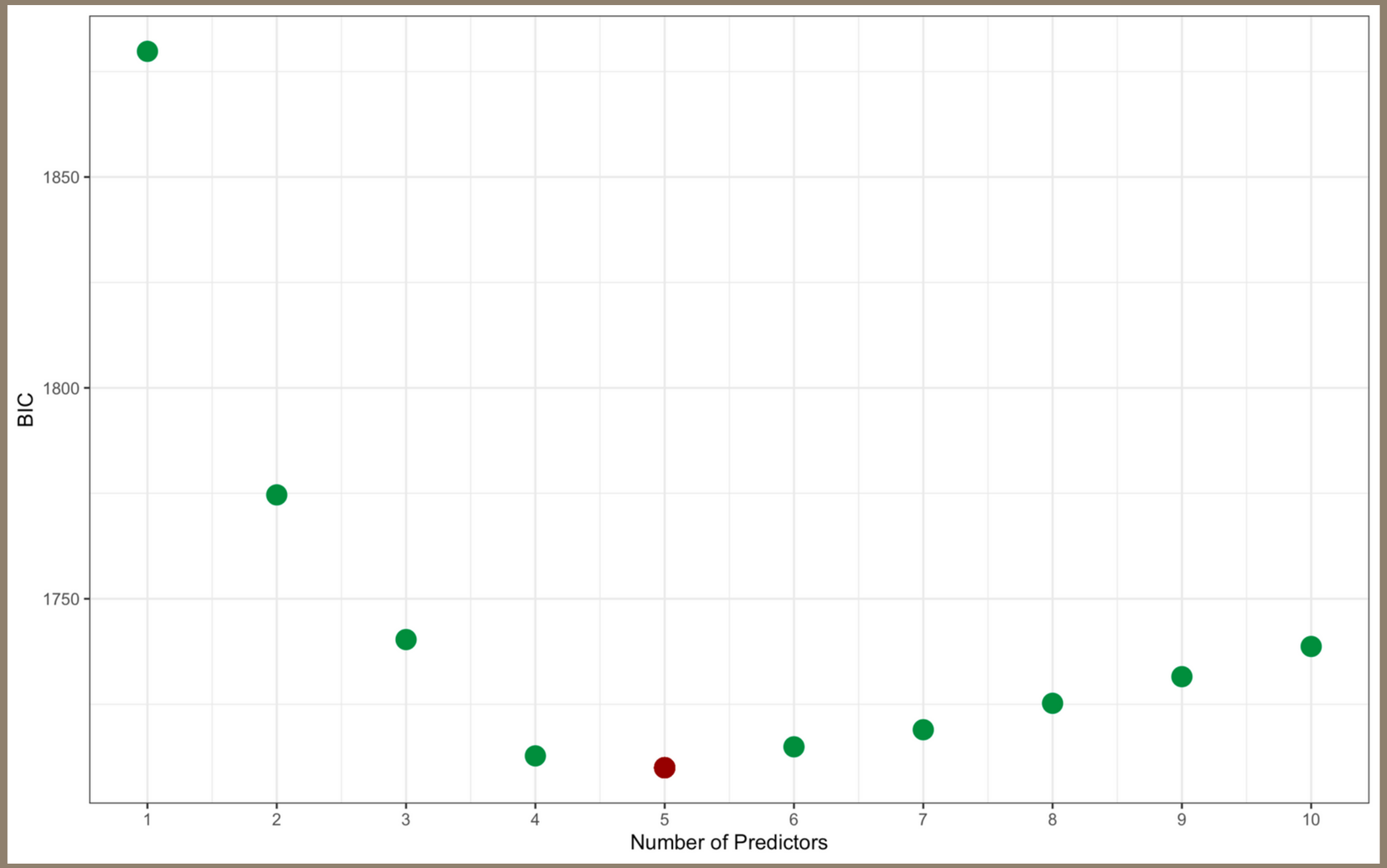
AIC



7 predictors

- volatile acidity
- chlorides
- sulphates
- alcohol
- bound sulphur dioxide
- citric acid
- pH

BIC



5 predictors

- volatile acidity
- chlorides
- sulphates
- alcohol
- bound sulphur dioxide

Model Selection

Criterion-based test:

- AIC
- BIC

Testing-based procedures:

- Backward elimination

To compare the models:

- ANOVA with χ^2 test

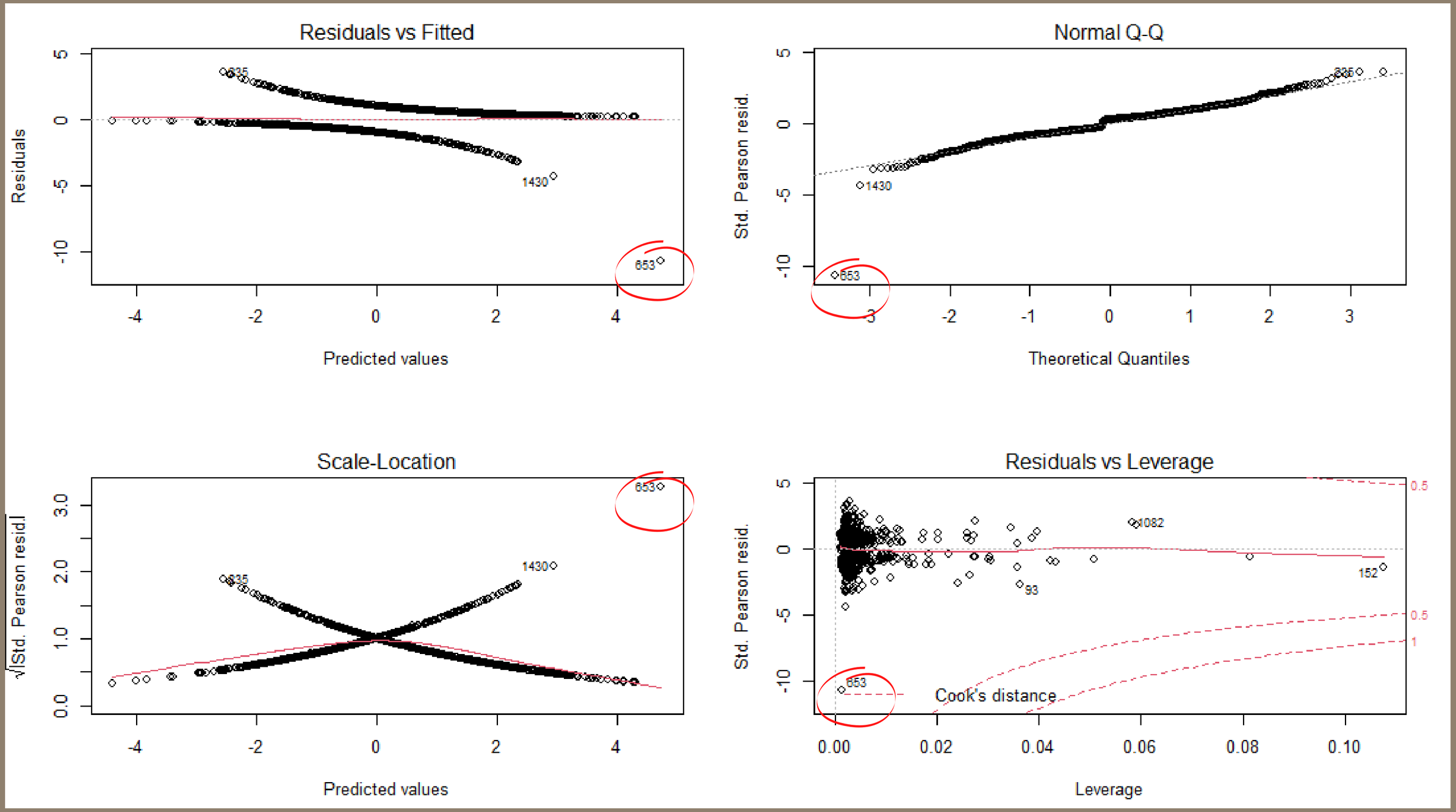
p-value: 0.1487

Goodness of fit:

- Hosmer-Lemeshow test

p-value: 0.7571

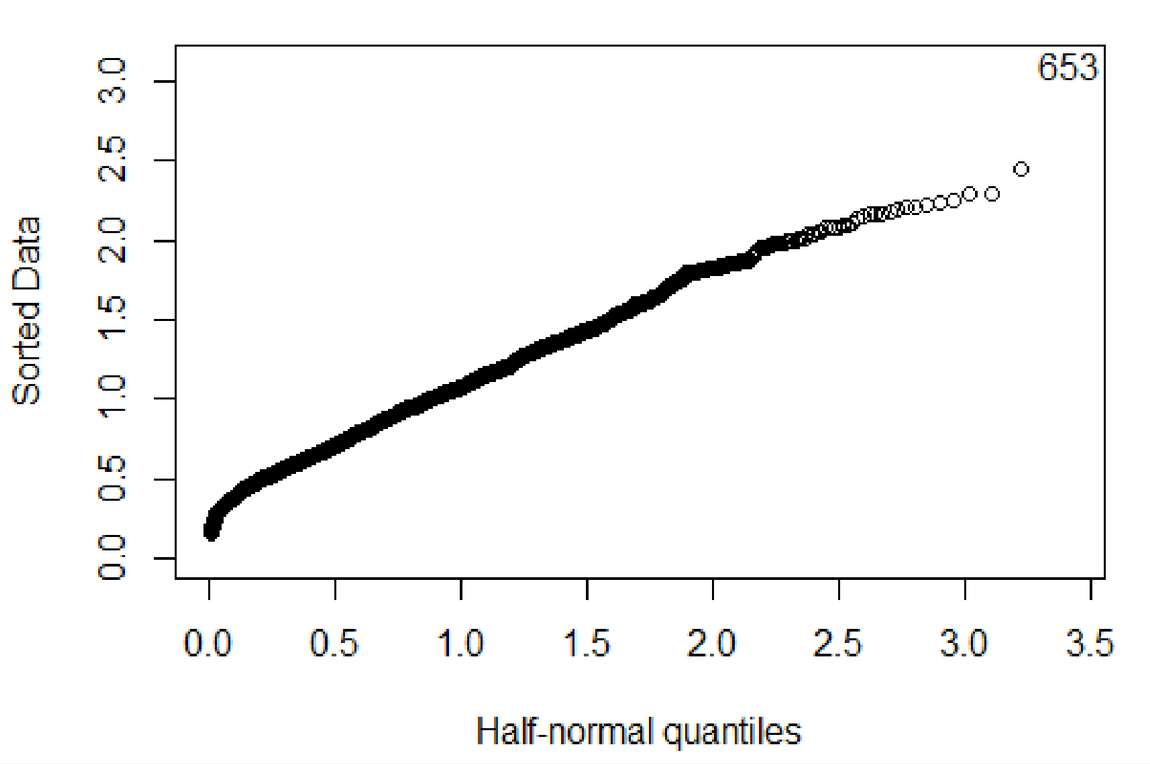
$$\begin{cases} H_0: Model\ suits\ the\ data \\ H_1: Model\ doesn't\ suit\ the\ data \end{cases}$$


Residuals vs Fitted
Residuals
Predicted values
635
1430
653
Normal Q-Q
Std. Pearson resid.
Theoretical Quantiles
635
1430
653
Scale-Location
√|Std. Pearson resid.|
Predicted values
653
1430
635
Residuals vs Leverage
Std. Pearson resid.
Leverage
Cook's distance
1082
152
93
653
0.5
0.5
1

Model Diagnostic

Checking for outliers

Studentized residuals in halfnorm plot:



Test:

$$|t_i| > t_{\alpha/(2*n)}(n-p-1)$$

$$t_{\alpha/(2*n)}(n-p-1) = 4.17604$$

$$\begin{cases} H_0 : i\text{-th observation is not an outlier} \\ H_1 : i\text{-th observation is an outlier} \end{cases}$$

Model Diagnostic

Checking for influential observations

- Cook's distance
- dfbeta function observation n° 755

summary estimates with 755

```
Coefficients:
                    Estimate Pr(>|z|)
(Intercept)         -8.09506  < 2e-16 ***
volatile.acidity    -2.91314 3.66e-15 ***
chlorides           -4.45501  0.00186 **
alcohol              0.86127  < 2e-16 ***
sulphates            2.71473 2.43e-10 ***
bound.sulfur.dioxide -0.01639 1.84e-11 ***
```

summary estimates without 755

```
Coefficients:
                    Estimate Pr(>|z|)
(Intercept)         -8.07995  < 2e-16 ***
volatile.acidity    -2.90497 4.47e-15 ***
chlorides           -4.84960 0.000959 ***
alcohol              0.86143  < 2e-16 ***
sulphates            2.72851 2.11e-10 ***
bound.sulfur.dioxide -0.01632 2.23e-11 ***
```

To Sum Up:

- Data visualisation
- Data manipulation
- Model selection
- Model diagnostic

CONCLUSIONS

BINOMIAL MODEL:

	Estimate
(Intercept)	-8.09506
volatile.acidity	-2.91314
chlorides	-4.45501
alcohol	0.86127
sulphates	2.71473
bound.sulfur.dioxide	-0.01639

MULTINOMIAL MODEL:

Coefficients:

	Value
volatile.acidity	-2.80816
chlorides	-5.91108
pH	-1.04281
sulphates	3.01230
alcohol	0.90718
bound.sulfur.dioxide	-0.01761

Intercepts:

	Value	Std. Error
bad|medium	5.1728	1.3022
medium|good	8.0583	1.3144

Conclusions

Initial goal: finding how physiochemical properties affect the quality of the Portuguese "Vinho Verde"

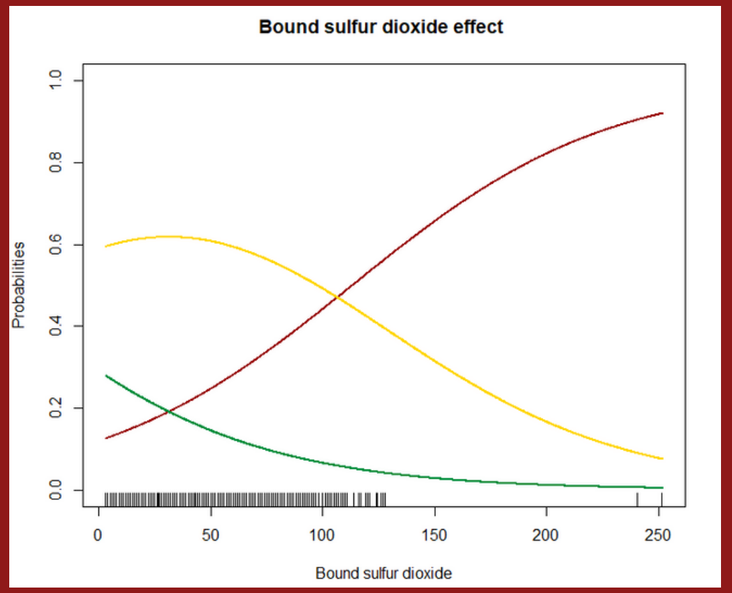
The two different approaches that we used lead us to similar models

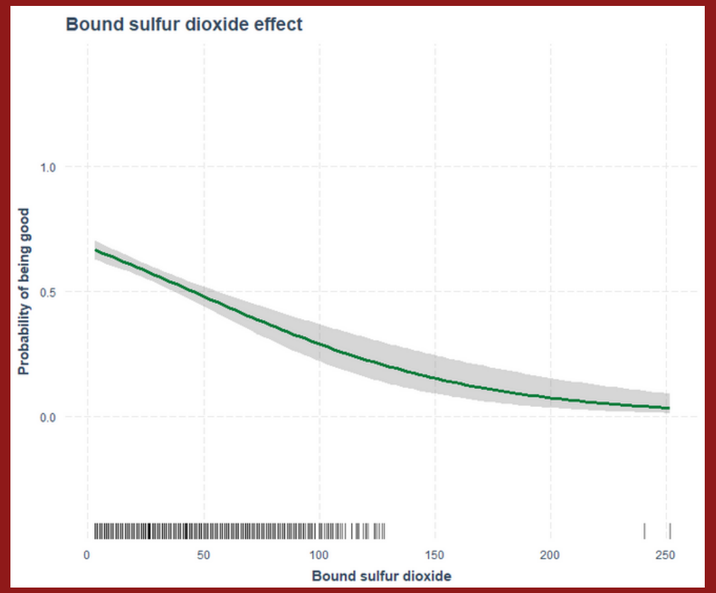
- Same 5 variables significant in the 2 models
- The only different variable is pH
- The effects of the common variables have the same sign

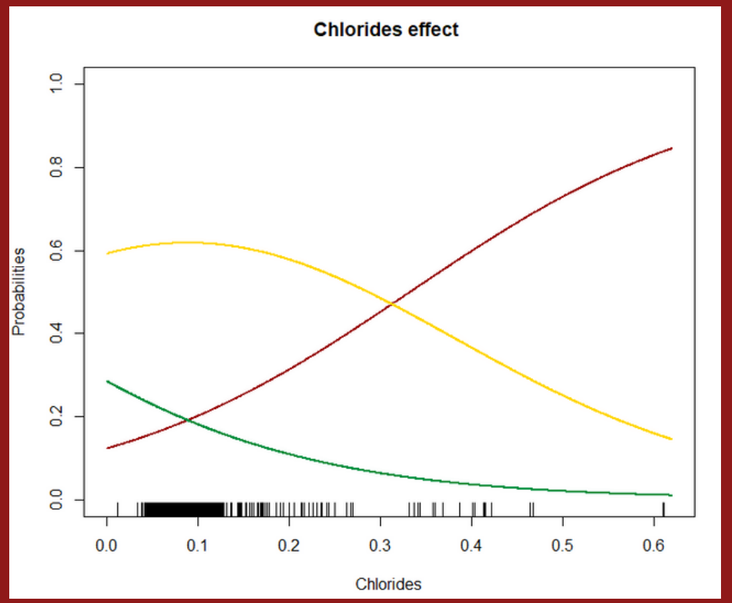
Multinomial

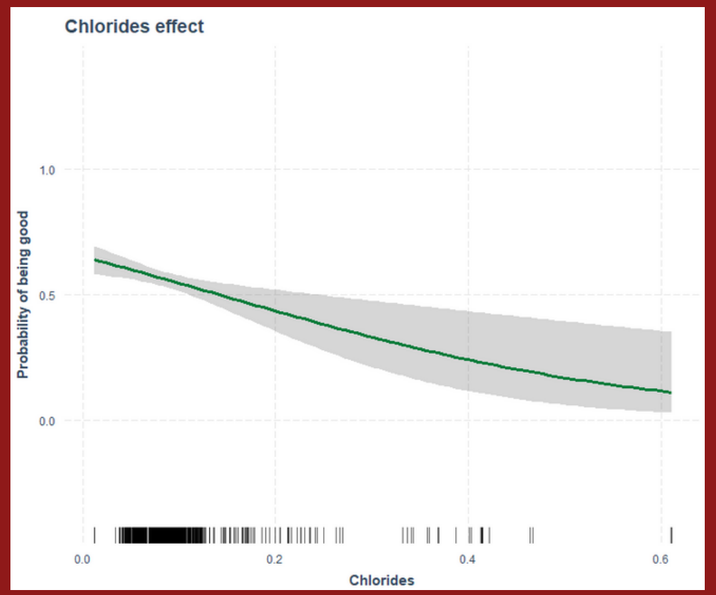
Binomial

bad
medium
good





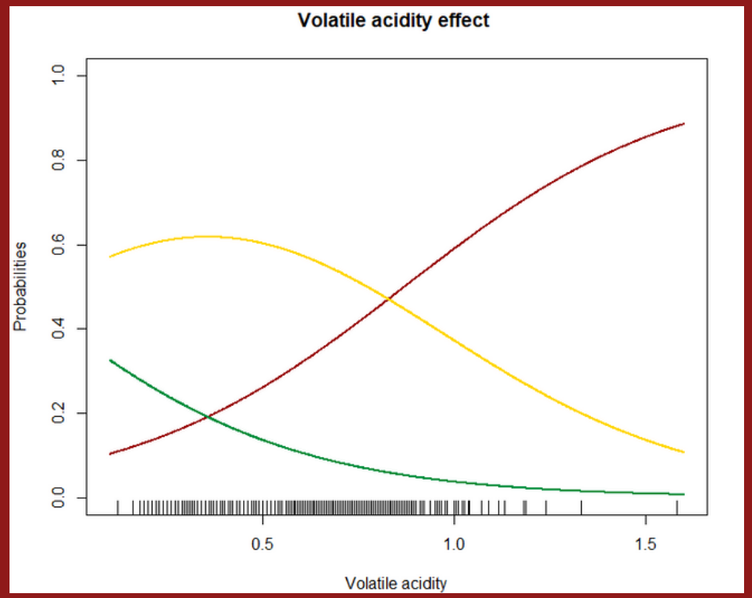


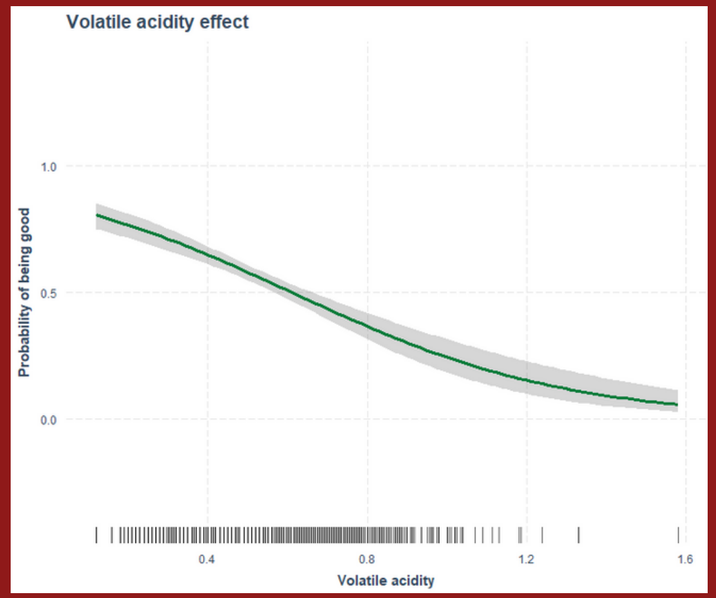


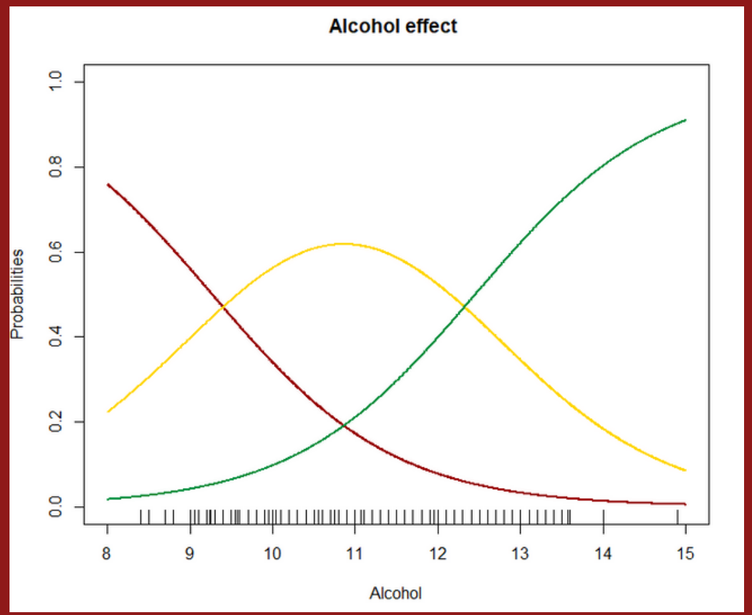
Multinomial

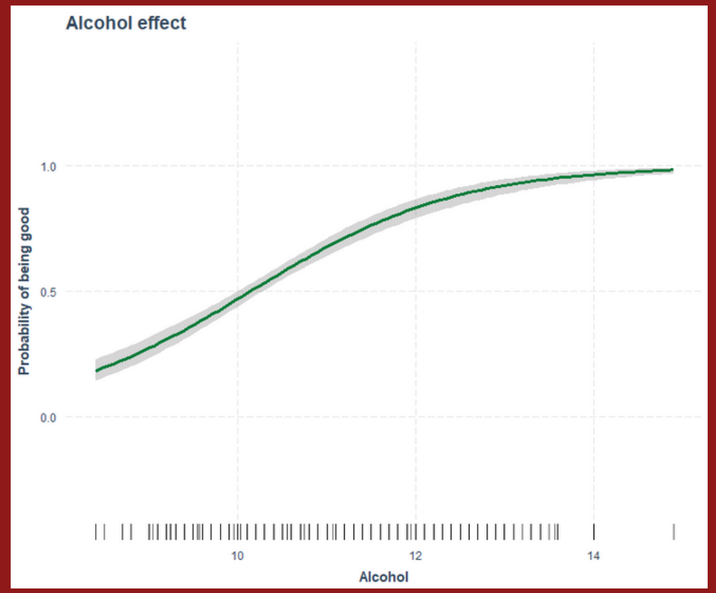
Binomial

bad
medium
good









Multinomial

Binomial

bad
medium
good