Bayesian Hierarchical Modeling On Wine Data
Jiaqi Sun (sunjiaqi)
December 20, 2024

Abstract

This project explores the use of Bayesian hierarchical regression to analyze wine quality, finding relations between chemical properties and taste scores. Bayesian methods enable efficient learning across wine types, addressing challenges like unequal sample sizes and multicollinearity. Bayes factors model selection is employed to identify the most relevant predictors. Posterior predictive checks confirm excellent model fit, and the results align with frequentist mixed-effects models. The model identifies key predictors of quality, such as alcohol content, residual sugar, and volatile acidity. This study demonstrate the value of hierarchical models in capturing group-level heterogeneity and highlight the important factors in enhancing wine quality.

1 Introduction

Wine has garnered increasing attention due to its antioxidant properties from grapes, with the industry's annual revenue projected to reach $380 billion by 2029 [4]. The evaluation of wine quality is based on two factors: physicochemical properties and sensory assessments conducted by experts. The physicochemical properties, such as pH, alcohol content, are straightforward to measure. However, human taste—a critical aspect of evaluation—has long been challenging to quantify and analyze. This study aims to bridge the gap between human sensory perception and the physicochemical characteristics of wine by analyzing wine quality. Understanding this relationship not only provides valuable insights for wineries to produce higher-quality wines but also advances our knowledge of human taste.

Previous researches [3] [1] on wine quality has applied frequentist machine learning methods, such as neural networks and support vector machines. However, these methods lack a probabilistic framework for quantifying uncertainty. Some studies [2] [5] have used Bayesian multinomial logistic regression and variable selection techniques like PSIS-LOO, but these typically focus on red wine alone, without leveraging the structure and information between red and white wine groups.

To address this limitation, I employ a Bayesian hierarchical regression model to jointly analyze the quality characteristics of red and white wines. This hierarchical approach allows for efficient learning across groups by sharing information between them. Model selection is conducted using Bayes factors, and prior predictive and posterior predictive checks are performed to evaluate the model's validity and fit. Additionally, I compare the Bayesian hierarchical regression model with its frequentist counterpart, the mixed-effect hierarchical model, to assess their relative performance.

2 Dataset

For this analysis, I utilized the Wine Quality dataset from the UCI Machine Learning Repository, which includes chemical and physical measurements of red and white wines. Each wine sample is evaluated for quality by professional tasters, providing a comprehensive resource for modeling wine quality based on physicochemical tests. The dataset includes 11 independent variables, such as fixed acidity, residual sugar. Summary statistics of these variables are presented in Table 1 and 2. After removing duplicates, the dataset consists of 1359 observations for red wine and 3961 for white wine.

Exploratory data analysis reveals notable differences between red and white wines. Red wines generally have higher fixed acidity and volatile acidity, indicating greater acetic acid levels. In contrast, white wines contain significantly more residual sugar, making them sweeter overall. The variability in sugar content is also more pronounced in white wines. While the pH and density of the two wine types are similar, red wines exhibit a slightly broader pH range. In terms of quality ratings (Figure 1), red wines tend to cluster around lower scores, with quality 5 and 6 accounting for the majority of samples. White wines are more frequently rated at 6, suggesting that white wines may generally achieve higher quality ratings. Correlation analysis (Figure 2) further highlights key relationships among the variables. For instance, alcohol content shows a positive correlation with quality in both wine types. In red wines, fixed acidity correlates strongly with density, while volatile acidity is negatively associated with quality, indicating that excessive acetic acid detracts from taste. For white wines, a strong correlation between residual sugar and density is observed, consistent with the expected physical properties.

3 Methodology

The dataset consists of grouped data for red and white wines, with white wine samples outnumbering red by three. This imbalance poses challenges for traditional frequentist models, which may lead to significant bias and inefficient use of the data. To address this, I employed a Bayesian Hierarchical Regression Model, allowing for efficient information sharing across groups and mitigating the effects of data imbalance. In this analysis, wine quality is treated as a continuous score assumed to follow a normal distribution. Let $j = 1, 2$ represent the wine type. The Bayesian hierarchical regression model is specified as follows:

$$y_{ij} = \beta_0 + u_j + \beta^T x_{ij} + \epsilon_{ij}, \qquad j \in \{1, 2\}, \tag{1}$$

where y_{ij} is the quality score of the i-th wine in the j-th group, x_{ij} is a vector contains the predictor variables, β_0 is the fixed intercept, and β is a vector of the fixed effect coefficients. u_j and ϵ_{ij} are the random effects accounting for variability at group and individual level. The fixed-effect parameters represent population-level trends, while the random effect u_j captures group-specific deviations. This hierarchical structure enables information sharing across groups, enhancing model flexibility and accuracy.

Predictors were normalized to accelerate convergence, ensure comparability across variables and simplify the selection of priors. A weakly informative $\text{Normal}(5, 2)$ prior was assigned to the intercept, centered around the mid-range of the quality scores. Similarly, coefficients were given $\text{Normal}(0, 1)$ priors. For group-level random effects, a $\text{Cauchy}(0, 0.5)$ prior was employed to balance flexibility with regularization. These priors were chosen to ensure the prior predictive distribution covers the observed range of quality scores. Prior predictive check was also conducted.

The dataset includes several predictors that describe related properties, such as different acidity measures, which may contribute redundantly to the response. Including all such predictors risks introducing multicollinearity. Model selection was performed using Bayes Factor (BF), a statistical measure that evaluates the relative evidence for competing models. By sequentially comparing a full model with reduced models that excluded one predictor at a time, I identified a subset of variables that optimized predictive accuracy. Posterior predictive checks were conducted to assess the model's ability to replicate observed data. To provide a baseline comparison, I also fitted a frequentist mixed-effects model, which shares a hierarchical structure with the Bayesian approach but relies on point estimates.

4 Results

The prior predictive check (Figure 3) demonstrate that the priors effectively covers the range of the quality variable. Though it slightly exceeds the observed range of quality scores, this is acceptable as long as the distribution adequately encompasses the data. Model comparison based on Bayes Factors (Table 3) indicates that the predictors chlorides and citric acid contribute minimally to the model, with Bayes Factors of 0.43 and 0.03, respectively. The remaining predictors show very large Bayes Factors (all greater than 100), signifying their substantial contribution to explaining wine quality.

The histograms and trace plots (Figure 5) of the posterior samples illustrate excellent convergence of the simulations. The trace plots for each parameter exhibit substantial overlap across chains with no discernible trends, indicating good mix. Additionally, all parameter estimates have ‘Rhat‘ values equal to 1, and the Effective Sample Sizes (ESS) exceed 2000 out of 6000 simulations, confirming the reliability of the posterior estimates. All predictors retained in the final model exhibit significant effects on wine quality, as indicated by their 95% credible intervals not crossing zero. Predictors such as volatile acidity and density negatively impact wine quality, while others, including alcohol and residual sugar, show positive effects. Posterior predictive checks (Figure 4) provide further evidence of the model’s strong performance. The scatter plot of the posterior mean and standard deviation indicates the model accurately captures the key statistical characteristics of the data. The posterior predictive density closely aligns with the histograms of the real data, demonstrating excellent fit both at the dataset level and within each wine group.

The comparison between the Bayesian hierarchical regression model and the mixed-effects model is summarized in Table 4. The parameter estimates from the Bayesian model closely match those from the frequentist model due to the large dataset size and the use of weakly informative priors. While Bayesian models provide a more flexible framework for incorporating prior knowledge and handling smaller sample sizes, frequentist approaches are computationally faster. Both methods achieve comparable performance, but Bayesian approaches offer advantages in scenarios where prior information is critical or data is limited.

5 Conclusion & Future Study

This study highlights the relationship between wine quality and key physicochemical attributes. Factors such as alcohol and residual sugar have the strongest positive impacts on wine quality, while volatile acidity and density negatively affect quality. These findings suggest that wineries could prioritize optimizing alcohol content and residual sugar levels while minimizing volatile acidity to enhance the perceived quality of their wines.

However, the study has limitations. The model is simplified and does not account for interactions between predictors. Incorporating random slope would improve accuracy but increase model complexity. Additionally, the dataset lacks external factors like grape origin and climate. Future work could expand the dataset to include these variables for a more comprehensive understanding of wine quality.

References

[1] Paulo Cortez et al. “Modeling wine preferences by data mining from physicochemical properties”. In: *Decision Support Systems* 47.4 (2009). Smart Business Networks: Concepts and Empirical Evidence, pp. 547–553. ISSN: 0167-9236. DOI: `https://doi.org/10.1016/j.dss.2009.05.016`. URL: `https://www.sciencedirect.com/science/article/pii/S0167923609001377`.

[2] X. Meng. *Bayesian Analysis for Red Wine Quality and Prediction*. Accessed: 2024-12-18. 2019. URL: `https://st540.wordpress.ncsu.edu/files/2019/05/Meng.pdf?`.

[3] Matteo Spronesti. *Wine Quality Classification*. Accessed: 2024-12-18. 2024. URL: `https://github.com/mspronesti/bayesian-wine-quality`.

[4] Statista. *Global Wine Market Size*. Accessed: 2024-12-18. 2024. URL: `https://www.statista.com/statistics/922403/global-wine-market-size/`.

[5] Aki Vehtari. *Bayesian variable selection for red wine quality ranking data*. Accessed: 2024-12-18. 2024. URL: `https://avehtari.github.io/modelselection/winequality-red.html`.

Tables & Figures

Figure 1: The distribution of wine quality for each type of wine.

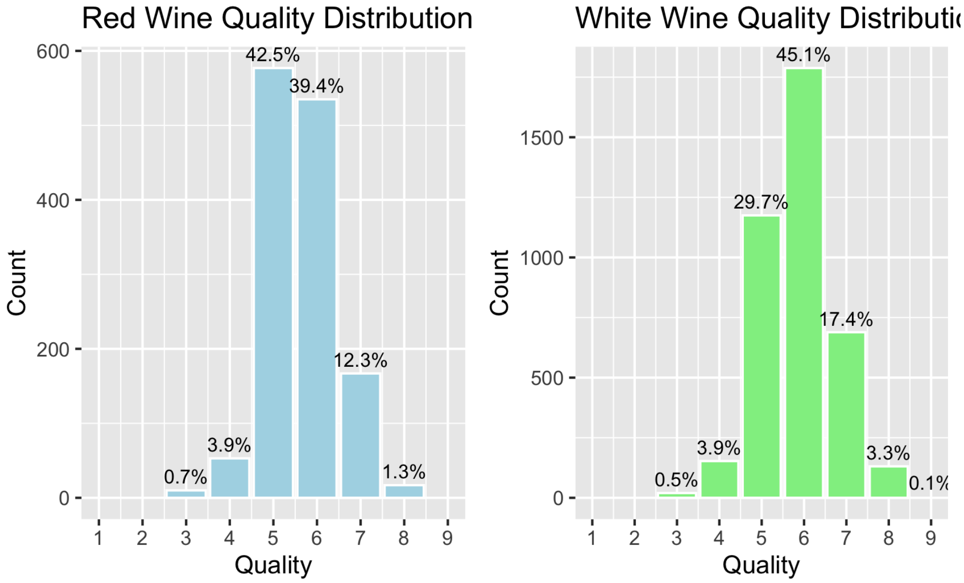

Figure 2: The heat map for correlations between variables

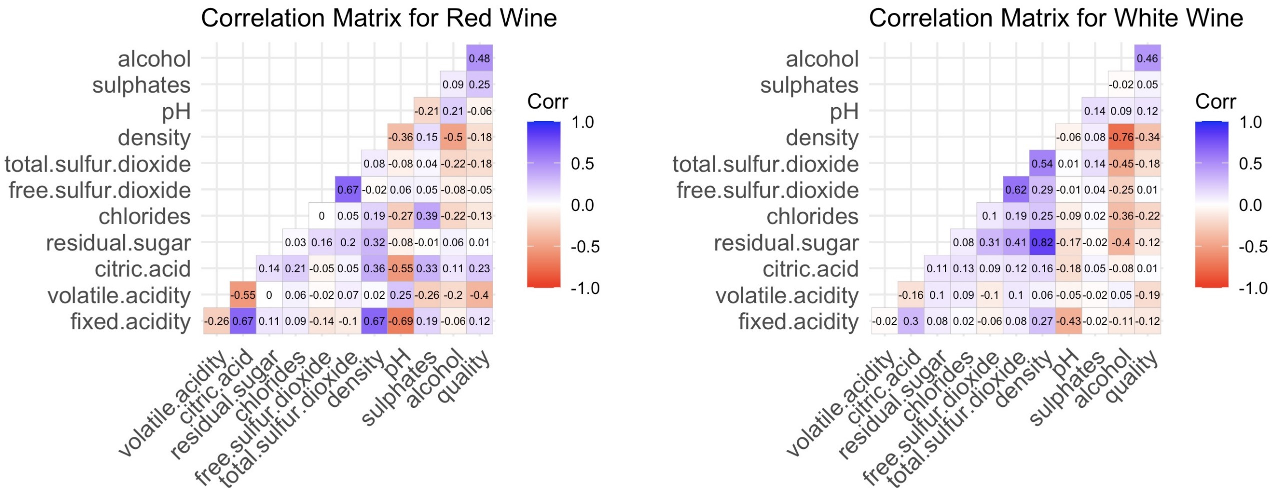

Figure 3: The prior predictive model checking. It compares the prior predictive density with real data.

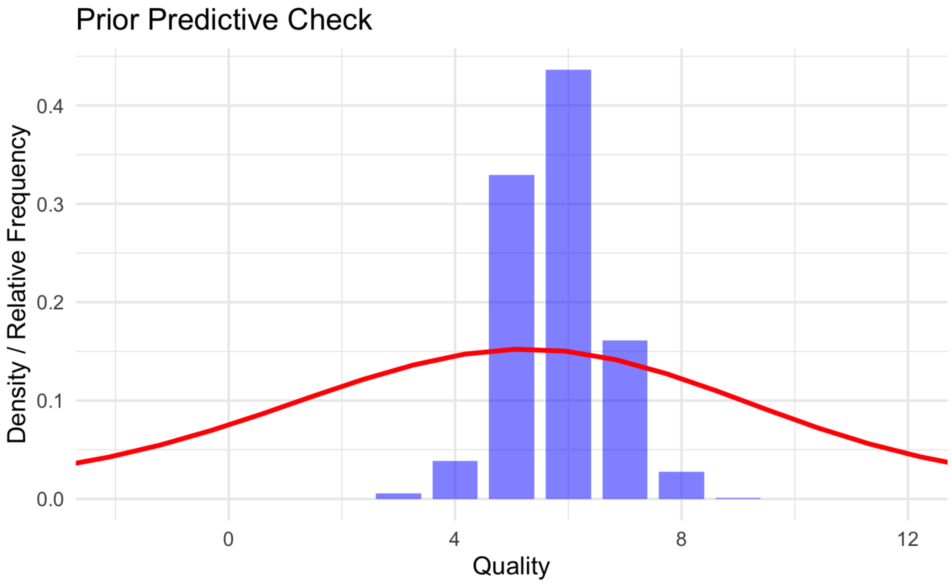

Figure 4: The posterior predictive model checking. Figure (a) checks two summary statistics of the posterior predictive samples. Figure (b) compares the posterior predictive density with the complete data histogram. Figure (c) and (d) compare the same setting but in different groups.

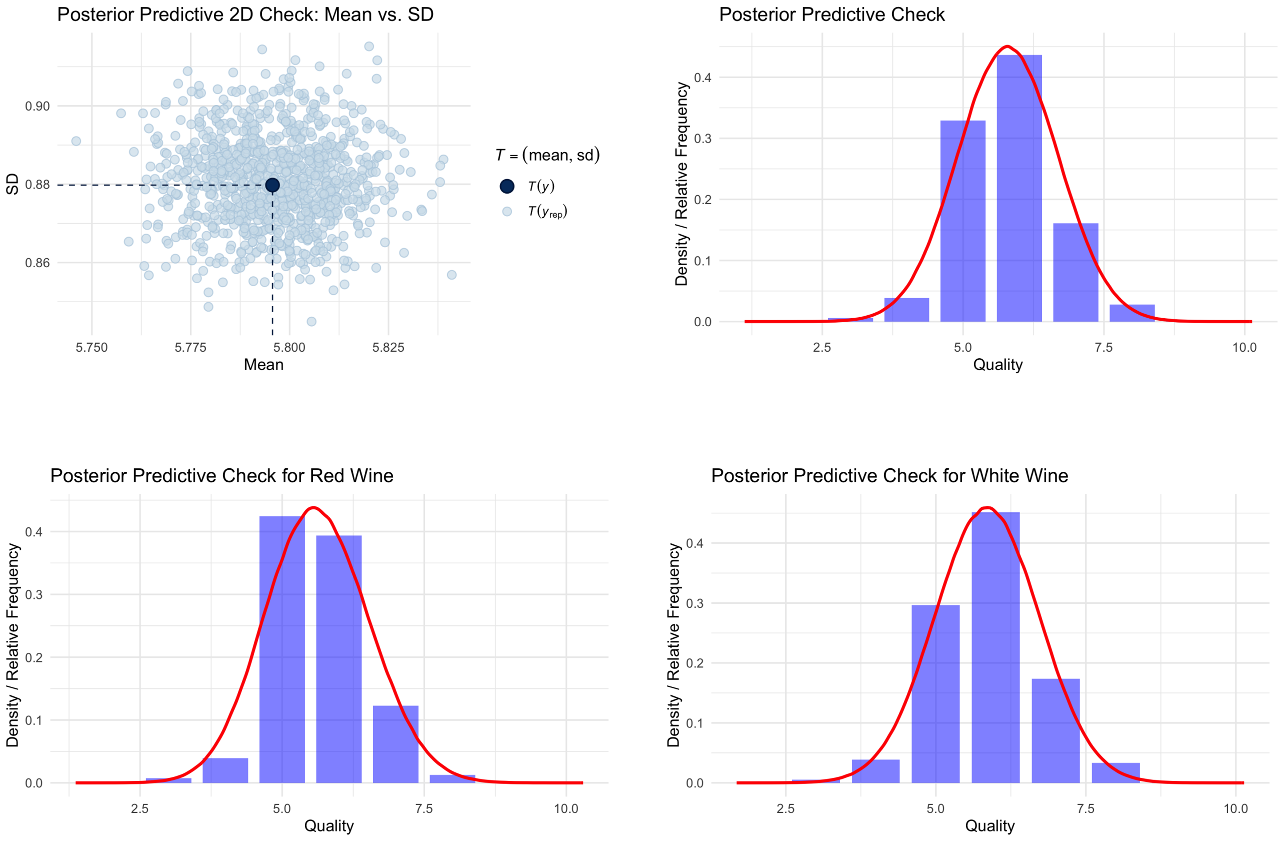

Figure 5: The histogram and trace plot for each random variable

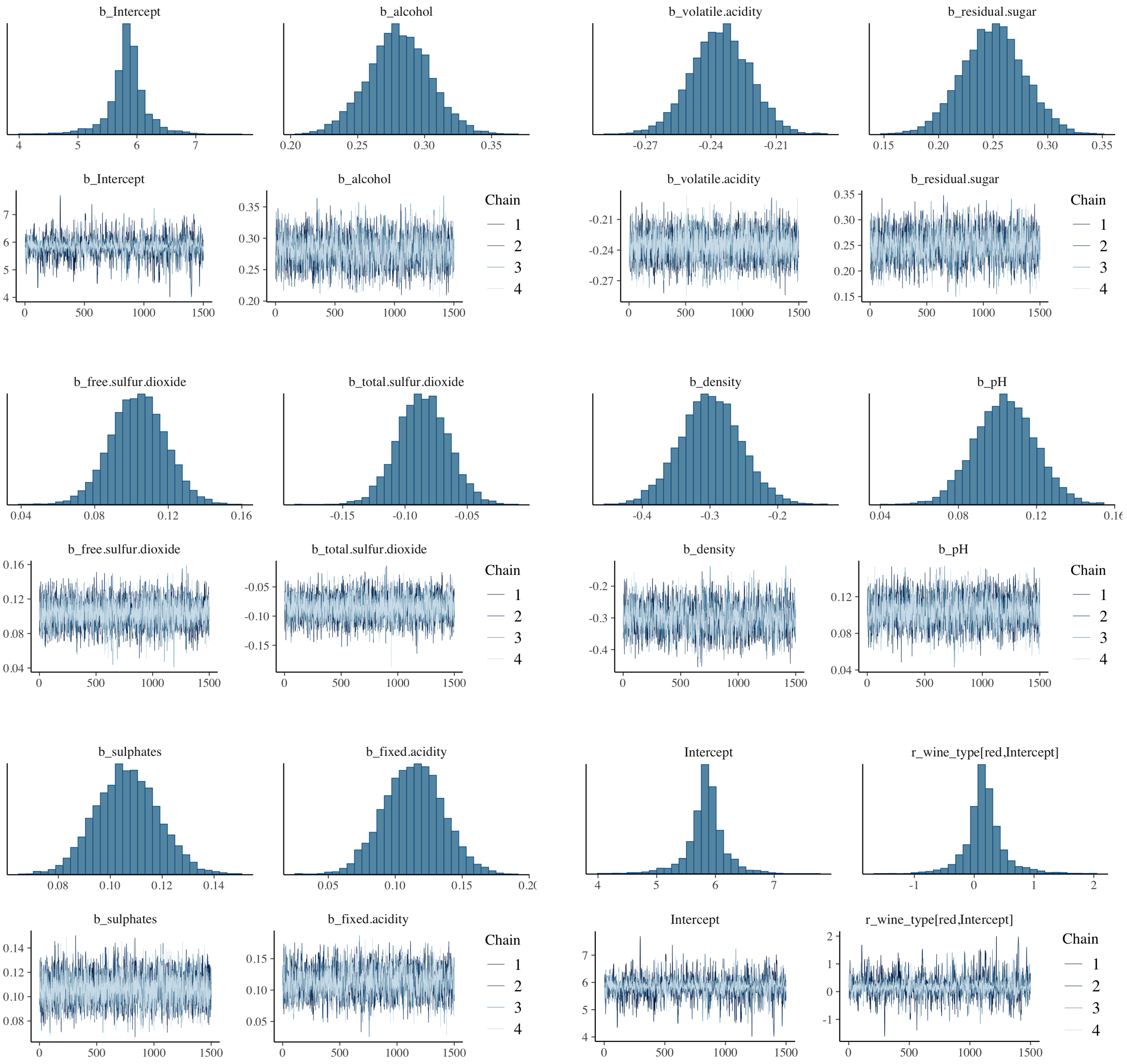

Table 1: Summary Statistics for Red Wine.

Attribute	Min	Max	Median	Mean	Standard Deviation
Fixed Acidity	4.60	15.90	7.90	8.31	1.7370
Volatile Acidity	0.12	1.58	0.52	0.53	0.1830
Citric Acid	0.00	1.00	0.26	0.27	0.1955
Residual Sugar	0.90	15.50	2.20	2.52	1.3523
Chlorides	0.012	0.611	0.079	0.088	0.0494
Free Sulfur Dioxide	1.00	72.00	14.00	15.89	10.4473
Total Sulfur Dioxide	6.00	289.00	38.00	46.83	33.4089
Density	0.9901	1.0037	0.9967	0.9967	0.0019
pH	2.740	4.010	3.31	3.31	0.1550
Sulphates	0.33	2.00	0.62	0.6587	0.1707
Alcohol	8.40	14.90	10.20	10.43	1.0821

Table 2: Summary Statistics for White Wine.

Attribute	Min	Max	Median	Mean	Standard Deviation
Fixed Acidity	3.80	14.20	6.80	6.84	0.8669
Volatile Acidity	0.08	1.10	0.26	0.28	0.1034
Citric Acid	0.00	1.66	0.32	0.33	0.1224
Residual Sugar	0.60	65.80	4.70	5.92	4.8616
Chlorides	0.009	0.346	0.042	0.046	0.0231
Free Sulfur Dioxide	2.00	289.00	33.00	34.89	17.2100
Total Sulfur Dioxide	9.00	440.00	133.00	137.20	43.1291
Density	0.9871	1.0390	0.9935	0.9938	0.0029
pH	2.720	3.820	3.180	3.195	0.1515
Sulphates	0.2200	1.0800	0.4800	0.4904	0.1135
Alcohol	8.00	14.20	10.40	10.59	1.2171

Table 3: Bayes Factor Comparison for Full Model Over Reduced Model

Predictor	**Bayes Factor**
Alcohol	3.90e+28
Volatile Acidity	2.72e+38
Residual Sugar	2.68e+13
Chlorides	4.29e-01
Free Sulfur Dioxide	4.88e+08
Total Sulfur Dioxide	4.47e+02
Density	2.32e+07
pH	4.08e+06
Sulphates	1.98e+15
Fixed Acidity	6.04e+02
Citric Acid	3.40e-02

Table 4: Comparison of Bayesian Hierarchical Regression and Mixed Effect Model Results. The two CI stand for credible interval and confidence interval.

Predictor	**Bayesian Estimate**	**Bayesian 95% CI**	**Bayesian Bulk_ESS**	**Mixed Effect Estimate**	**Mixed Effect 95% CI**
Intercept	5.84	[5.07, 6.51]	1656	5.87	[5.59, 6.14]
Alcohol	0.28	[0.24, 0.33]	2274	0.28	[0.24, 0.33]
Volatile Acidity	-0.24	[-0.26, -0.21]	4914	-0.24	[-0.26, -0.21]
Residual Sugar	0.25	[0.19, 0.31]	1978	0.25	[0.19, 0.31]
Free Sulfur Dioxide	0.10	[0.08, 0.13]	4467	0.11	[0.08, 0.13]
Total Sulfur Dioxide	-0.09	[-0.13, -0.05]	4087	-0.09	[-0.13, -0.05]
Density	-0.30	[-0.39, -0.21]	1940	-0.30	[-0.39, -0.21]
pH	0.10	[0.07, 0.13]	2775	0.10	[0.07, 0.13]
Sulphates	0.11	[0.08, 0.13]	5074	0.11	[0.08, 0.13]
Fixed Acidity	0.11	[0.07, 0.16]	2314	0.11	[0.07, 0.16]