Physicochemical Predictors of Wine Quality: A Bayesian Analysis

Tristan Mesaros

Georgia Institute of Technology

April 20, 2025

Abstract

In this analysis, Bayesian linear regression is applied to a subset of the UCI Red Wine Quality dataset to investigate how various physicochemical attributes influence perceived wine quality. Two competing models are specified: the first employs weakly informative normal priors, while the second incorporates Laplace priors for most predictors, with uniform priors placed on pH and alcohol to reflect domain-informed constraints. Model comparison is conducted using Leave-One-Out (LOO) cross-validation to assess out-of-sample predictive performance. The model with superior predictive accuracy is then used to solve a constrained optimization problem aimed at identifying the combination of feature values that maximizes the predicted wine quality, specifically targeting a quality score of 10. Posterior predictive checks are then used to assess the credibility of this estimate. Overall, the analysis demonstrates how Bayesian tools can be used not just for modeling but also for making informed predictions in a practical context.

1 Introduction

Wine production and consumption have long held significant cultural, economic, and historical importance in many countries, particularly in France, where wine is regarded as both a national symbol and a staple of everyday life. As a French student, I have a personal connection to wine culture, which inspired me to explore this topic through the lens of Bayesian statistical analysis. Wine is a complex product whose quality depends on a wide range of physicochemical factors, from acidity and sugar content to alcohol concentration and pH balance. The data set used in this study originates from a well-known research effort by Cortez et al. (2009), which analyzed the physicochemical properties of red wine samples alongside the quality scores assigned by certified sommeliers. Modeling the relationship between these physicochemical properties and wine quality is not only of academic interest but also has significant implications for the wine industry. Identifying the most influential chemical markers can support quality control, guide production processes, and even inform product development. However, the data-generating process in this context is inherently noisy, and the predictors often exhibit Multicollinearity, making it challenging to derive robust and interpretable models using classical statistical techniques alone. To address these challenges, this study adopts a Bayesian modeling framework, which offers several key advantages. First, Bayesian methods allow for the incorporation of prior knowledge or domain expertise into the modeling process. Second, the Bayesian approach provides a coherent and principled way to quantify uncertainty in parameter estimates and predictions. Finally, Bayesian model comparison techniques, such as Leave-One-Out cross-validation (LOO), offer a robust method for evaluating model performance without overfitting. By comparing models with different prior structures, including weakly informative normal priors and a combination of Laplace and uniform priors, this analysis seeks not only to understand the determinants of wine quality, but also to leverage these insights for predictive optimization.

2 Data Processing

Before diving into Bayesian analysis, I randomly selected a subset of 150 observations. There were a couple reasons for this: first, it helped reduce computation time during sampling, and second—and more importantly—Bayesian methods are especially powerful with smaller datasets. One of the advantages of Bayesian inference is that it can still provide rich, interpretable results even when data is limited, unlike many frequentist approaches that need large samples to perform well.

Before building the models, I standardized all the predictor variables so they'd be on the same scale. This helps the sampler run more efficiently and also makes it easier to compare the influence of different variables. I also saved the means, standard deviations, and min/max values so I could later translate predictions back into their original units. This way, even though the models work on scaled data, the results still make sense in real-world terms.

3 Exploratory Analysis

To get a better sense of how the different wine characteristics relate to quality, I started by creating a correlation matrix for all the variables in the dataset. A few relationships stood out right away. Alcohol content showed a moderate positive correlation with quality (0.50), which makes sense since higher alcohol levels are often associated with fuller-bodied wines that tend to score better in sensory evaluations. Sulphates also had a moderate positive correlation with quality (0.30), while volatile acidity showed a moderate negative correlation (-0.43), likely because higher levels of volatile acidity are associated with sour, vinegary flavors that negatively affect a wine's perceived quality.

Most of the other variables, like residual sugar, density, chlorides, and pH, had fairly weak linear relationships with quality, with correlations typically hovering between -0.20 and 0.20. There were also a couple of expected patterns within the wine's chemical properties themselves, such as a strong negative correlation between fixed acidity and pH (-0.71), and a strong positive correlation between free and total sulfur dioxide (0.64). These initial exploratory results helped narrow down which predictors to prioritize when building the Bayesian regression models, focusing on those variables that showed stronger associations with quality. The full correlation matrix is shown in Figure 1.

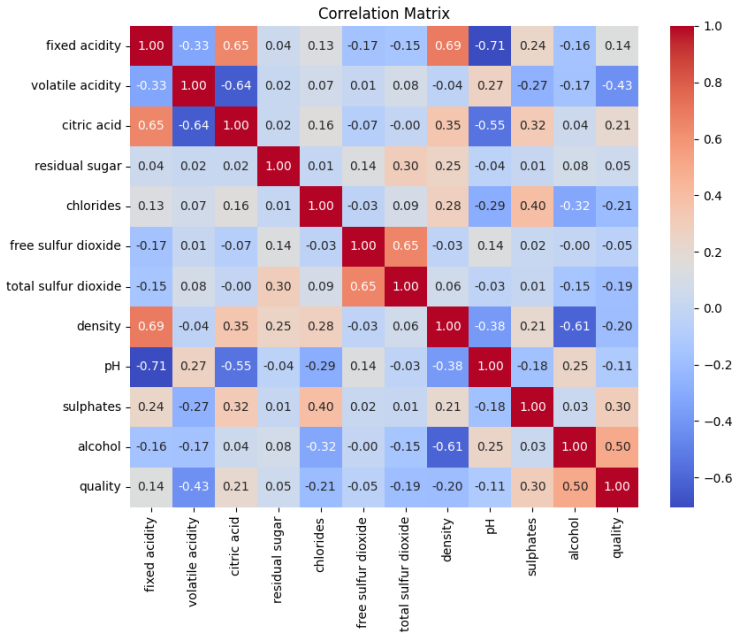

Figure 1: Correlation matrix of physicochemical parameters in wine making.

4 Bayesian Methodology

4.1 Statistical Approach and Models

This analysis uses a Bayesian regression approach to explore how different physicochemical properties of wine are associated with its quality rating. The goal is to estimate the strength and direction of these relationships while accounting for uncertainty in the data and model parameters. Bayesian methods are particularly useful in this context because they provide full posterior distributions for each parameter, rather than single point estimates, which makes it easier to quantify uncertainty and incorporate prior beliefs when appropriate. The outcome variable in this study is the wine's quality score, treated as continuous for modeling purposes. The regression model assumes the following likelihood:

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij}$$

where μ_i is the predicted quality for wine i, β_0 is the intercept, β_j are the regression coefficients for each predictor, and σ represents the residual standard deviation. To investigate how different prior assumptions might influence the model results, two versions of the model were fit using different prior distributions on the regression coefficients.

4.1.1 Model 1: Weakly Informative Priors

In the first model, weakly informative priors were placed on the intercept and coefficients. Specifically, both the intercept and the regression coefficients were given Normal distributions centered at zero with a standard deviation of 1:

$$\beta_0, \beta_j \sim \text{Normal}(0, 1)$$

The residual standard deviation, σ, was given a half-normal prior:

$$\sigma \sim \text{HalfCauchy}(1)$$

The idea behind using weakly informative priors is to regularize the estimates slightly while still allowing the observed data to primarily influence the posterior distributions. This is especially helpful when working with moderately sized datasets, as it can help prevent extreme or implausible estimates without being overly restrictive. Sampling was performed using the No-U-Turn Sampler (NUTS) from `PyMC`, with two chains and 2000 posterior draws per chain, following a 1000-sample tuning phase. Convergence diagnostics such as trace plots and R-hat values were examined to confirm that the sampler performed reliably.

4.1.2 Model 2: Laplace and Uniform Piors

The second model was designed to assess how sensitive the results might be to different prior choices. In this version, most of the regression coefficients were assigned Laplace

(double-exponential) priors, which are known to induce stronger shrinkage towards zero:

$$\beta_0, \beta_j \sim \text{Laplace}(0, 1)$$

For two predictors, domain-specific Uniform priors were applied. The coefficient for `pH` was assigned a Uniform prior between -2 and 2, reflecting the belief that its effect size is likely modest and constrained:

$$\beta_{\text{pH}} \sim \text{Uniform}(-2, 2)$$

Similarly, the coefficient for `alcohol` was assigned a Uniform prior between 0 and 3, under the assumption that higher alcohol levels tend to increase wine quality, and the effect is expected to be positive and moderately strong:

$$\beta_{\text{alcohol}} \sim \text{Uniform}(0, 3)$$

Unlike Model 1, where the intercept was given a weakly informative Normal$(0, 1)$ prior, Model 2 assigned the intercept a more informative prior centered around 5 with a standard deviation of 2:

$$\text{intercept} \sim \text{Normal}(5, 2)$$

This reflects prior knowledge about the typical range of wine quality ratings in the dataset, which are generally centered around 5 to 6. Since all predictors are standardized, this prior on the intercept essentially represents the expected average wine quality when all standardized predictors are at their mean value of zero. Including this more informative intercept prior in Model 2 allows for a fairer test of how different coefficient priors affect posterior estimates, while better aligning the model with known characteristics of the outcome variable. As with Model 1, the residual standard deviation was modeled with a Half-Cauchy prior:

$$\sigma \sim \text{HalfCauchy}(1)$$

Sampling and diagnostic procedures for this model were the same as for Model 1.

4.2 Model Comparison and Selection

To formally evaluate and compare the predictive performance of the two Bayesian regression models, Leave-One-Out Cross-Validation (LOO-CV) was employed. LOO-CV is a widely recommended model comparison technique in Bayesian analysis, providing an estimate of a model's expected predictive accuracy by sequentially excluding each observation from the dataset, refitting the model, and computing the log predictive density for the omitted observation. This approach offers a principled, out-of-sample estimate of predictive performance while accounting for model complexity and overfitting risk.

As shown in Table 1, the model incorporating Laplace and Uniform priors achieved a slightly higher expected log pointwise predictive density (ELPD_LOO = -147.003) than the model using weakly informative Normal priors (ELPD_LOO = -147.809). While

Table 1: Leave-One-Out Cross-Validation (LOO-CV) Results for Model Comparison

Model	ELPD_LOO	Rank
Laplace Priors	-147.003	0
Weak Priors	-147.809	1

the difference in ELPD values is modest, it suggests that the Laplace prior model offers marginally improved out-of-sample predictive performance. Given this result, alongside the appealing shrinkage properties of Laplace priors for regularizing coefficient estimates, the Laplace + Uniform prior model was selected as the preferred specification for inference and prediction in this analysis.

5 Results

The final analysis focused on the Laplace + Uniform prior model, which demonstrated slightly superior out-of-sample predictive performance based on Leave-One-Out Cross-Validation (LOO-CV) and exhibited desirable shrinkage behavior on less influential predictors.

5.1 Posterior Parameter Estimates

Table 2 summarizes the posterior means, standard deviations, and 95% highest density intervals (HDIs) for each parameter in the Laplace prior model. Among the physico-chemical properties considered, `alcohol` showed the strongest positive association with wine quality, with a posterior mean of 0.383 and a 95% HDI ranging from 0.196 to 0.576. This result suggests that, holding other predictors constant, higher alcohol content is associated with increased quality scores.

Table 2: Posterior Summary Statistics for Laplace & Uniform Prior Model

Parameter	Mean	SD	2.5% HDI	97.5% HDI	R-hat
Intercept	5.674	0.051	5.571	5.772	1.00
Fixed Acidity	-0.039	0.135	-0.325	0.211	1.00
Volatile Acidity	-0.268	0.071	-0.396	-0.128	1.00
Citric Acid	-0.177	0.089	-0.342	0.004	1.00
Residual Sugar	0.025	0.064	-0.101	0.147	1.00
Chlorides	-0.159	0.064	-0.284	-0.032	1.00
Free Sulfur Dioxide	0.045	0.067	-0.088	0.170	1.00
Total Sulfur Dioxide	-0.121	0.073	-0.261	0.022	1.00
Density	0.052	0.136	-0.209	0.323	1.00
pH	-0.230	0.091	-0.402	-0.046	1.00
Sulphates	0.241	0.060	0.124	0.360	1.00
Alcohol	0.383	0.098	0.196	0.576	1.00
Sigma	0.619	0.037	0.550	0.695	1.00

`Sulphates` also exhibited a notable positive association with quality, with a posterior mean of 0.241 (95% HDI: 0.124 to 0.360). Conversely, `volatile acidity` had a moderate negative relationship with wine quality, with a posterior mean of -0.268 and a 95% HDI excluding zero (-0.396 to -0.128). This is consistent with prior expectations, as higher volatile acidity typically introduces undesirable sourness in wine.

Several other predictors, such as `citric acid`, `chlorides`, `pH`, and `fixed acidity`, had posterior means closer to zero with HDIs overlapping zero, indicating weak or uncertain effects on wine quality after accounting for the other variables in the model. The shrinkage effect of the Laplace priors was apparent in these parameters, pulling uninformative coefficients closer to zero, which helps avoid overfitting and improves model parsimony.

5.2 Posterior Predictive Performance

To assess the model's practical utility, posterior predictive checks were conducted by identifying optimized predictor values expected to achieve a wine quality rating of 10. Table 3 presents these estimated values, including both standardized values and their original-scale equivalents.

Table 3: Estimated Predictor Values to Achieve Wine Quality of 10 (Laplace Model)

Predictor	**Standardized Value**	**Original Scale**
Fixed Acidity	-1.751	5.000
Volatile Acidity	-1.989	0.180
Citric Acid	-1.331	0.000
Residual Sugar	5.289	8.600
Chlorides	-1.052	0.044
Free Sulfur Dioxide	3.905	52.000
Total Sulfur Dioxide	-1.114	10.000
Density	2.588	1.001
pH	-2.270	2.980
Sulphates	3.449	1.170
Alcohol	3.383	14.000

Using the posterior samples, the model predicted a mean wine quality of 9.89 for these optimized values, with a posterior standard deviation of 0.93 and a 95% credible interval ranging from 8.06 to 11.71. This result suggests that, under the model's assumptions, achieving a wine quality rating near the maximum observed value is possible with a specific combination of physicochemical properties. Notably, achieving this requires higher-than-average values for `alcohol`, `sulphates`, and `residual sugar`, along with lower values for `volatile acidity`, `chlorides`, and `total sulfur dioxide`.

Overall, the Laplace prior model provided interpretable and consistent estimates, while offering good predictive performance and robustness through the use of shrinkage priors.

6 Conclusion

This study demonstrates the utility of Bayesian linear regression as a robust and interpretable framework for modeling wine quality using the UCI Red Wine Quality dataset. By constructing and comparing two models—one employing weakly informative Normal priors and another integrating Laplace priors alongside domain-informed Uniform priors on alcohol and pH—we investigated the influence of prior specification on posterior inference and predictive performance. The model incorporating stronger prior structure exhibited marginally superior predictive accuracy, as indicated by leave-one-out cross-validation, underscoring the value of integrating substantive domain knowledge into the modeling process. Furthermore, the application of the posterior predictive distribution to identify optimal feature combinations for maximizing wine quality illustrates the potential of Bayesian methods not only for inference but also for forward-looking decision support. While the predicted probability of producing a wine rated 10 remains low, the analysis provides a principled, probabilistic framework for navigating such optimization tasks. Taken together, the results underscore the strengths of Bayesian modeling in yielding both interpretable estimates and actionable insights in applied settings.

References

[1] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). *Wine Quality*. UCI Machine Learning Repository. https://doi.org/10.24432/C56S3T.