Bayesian Inference on wine by location and variety

Candidate number: 50537

Contents

Introduction

I have used Bayesian inference to model how a wine is priced based on the following factors: country, province, and rating where province is a subcategory of country and the rating is assumed to be a valid measure of wine quality. Bayesian inference is interesting to apply here as wines from the same country or province are likely to have either similar methods, terroir (natural environment), or growing conditions (e.g. weather). This can also discern whether there are differences in value (how much price increases for an increase in quality) between countries. An attribute of Bayesian modelling which is helpful here is that for countries or provinces with less available data, they can use information from other countries or provinces to still make reasonably accurate predictions. For winemakers, this modelling can allow them to better compare themselves to others in their region to assess their pricing and value relative to their neighbouring competitors for a more strategic market position. My main research aim will be to see *how accurately and interpretably a hierarchical, Bayesian model can model a wine's price based on its country, province, and quality (rating) compared to frequentist models*. As well as this I can compare how old-world wines (such as France, Italy, and Greece) compare to new-world wines (such as the US, Australia, and Chile) and more emerging wine-making countries (such as England, and Canada) in terms of wine quality and wine value. There are also many other interesting features that can be looked into such as: "is price a good predictor of quality?", "how variable is wine quality and value between countries and provinces?" and "how much does location affect price, quality, and value?". I used Bayesian hierarchical models, as well as lasso and ridge regressions for comparison, to evaluate my research aim.

Dataset, Transformations and Data Exploration

The dataset I have chosen is `winemag-data-130k-v2.csv` (referred to in the code as just `wine`) Wine Reviews by 'zackthoutt', created in 2018, on Kaggle. The data was scraped from WineEnthusiast (a B Corp certified company that deals in wine reviews, shopping, education, storage and more) on 15th June 2017 and then again on 22nd November 2017. It has around 130,000 entries and includes data on: id (identification number), country, (tasting) description, designation (e.g. "reserve" or "classic"), points (rating), price (in USD), province, region 1 and region 2 (further details on location within the province), taster name and twitter handle, review title (including wine age), (grape) variety, and winery.

Firstly I removed columns based on the number of empty or missing values they contained which resulted in the removal of region 2, designation, twitter, taster name, and region 1 as they all had over twenty thousand missing values. For the removal of taster name, I will assume in my model it does not affect rating and that ratings are unbiased. I also removed id as it was unnecessary for my data analysis, and description, review title, and winery as they would have added too much complexity to the model making it less interpretable. During exploratory data analysis (EDA), I found that variety had 668 unique values and so I removed it to reduce model complexity and improve interpretability. Lastly, due to the fact that wine pricing can has extreme outlier values (see Table 2), I have chosen to transform price to log_price by taking the log (base 2) of price, in order to have a more linear relationship between the two which should be better for modelling purposes. Moreover, wine rating was scored 0-100 but all data are within 80 and 100 where a higher number means a better wine. For my analysis, I centered its distribution around zero by subtracting the mean rating from each rating, to create rating_c, with the aim of speeding up convergence. This resulted in choosing the following variables for my final model (see Table 1):

Variable name	Type of data	Description
`rating_c`	float (decimal)	The quality of the wine, based on the reviewer's opinion of it. Centered to have a mean of zero.
`log_price`	float (decimal)	Target variable. The log of the price of a bottle of the wine at the time of purchase for the review.
`country`	character (categorical)	The country the wine was produced in.
`province`	character (categorical)	The province (of the country) in which the wine was produced.

Table 1. *Variables*

The numeric data had the following properties (see Table 2):

variable name	***minimum***	***median***	***mean***	***maximum***
price	4.00	25.00	35.44	3300.00
log_price	2.00	4.64	4.78	11.69
rating	80.00	88.00	88.42	100.00
rating_c	-8.42	-0.42	0.00	11.58

Table 2. *Continuous Variables Summary (to 2 decimal places)*

The data was then cleaned by removing any rows with empty or null values, and removing provinces and countries which contained less than 10 wines as the level of data would be too small and therefore noisy (high variance) to get accurate results from, for my analysis. This resulted in a total of 120,245 entries that were appropriate for analysis. Also, I checked log_price to ensure that it does roughly follow a normal distribution (See Figure 1).

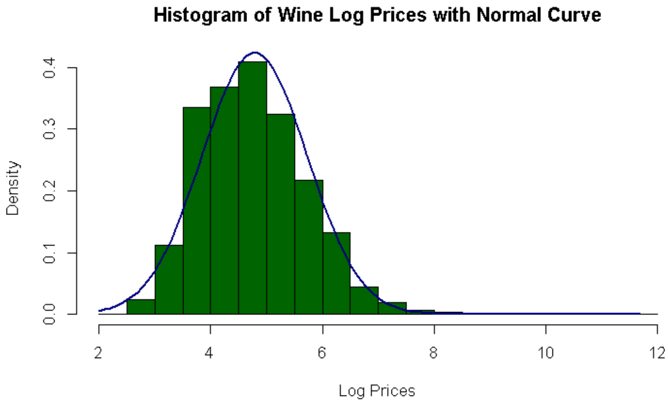

Figure 1. *Histogram of Log Wine Prices*

While exploring variation in prices between countries I found that it would be worth creating a hierarchical model as the prices (both on average and in terms of variance) clearly varied by country (see Figure 2).

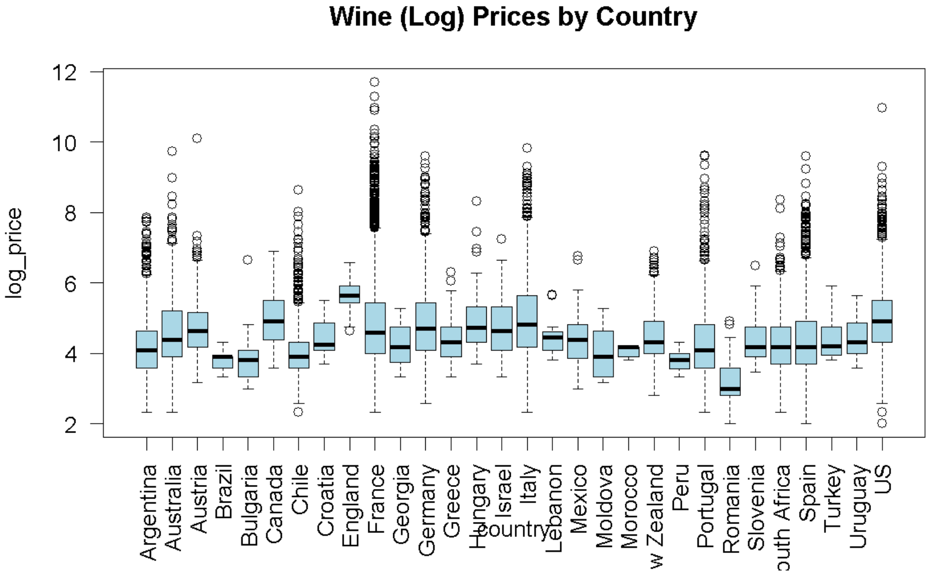

Figure 2*. Wine (Log) Prices by Country Boxplot*

Moreover, while exploring the data further, I found the prices seemed to vary substantially not only between countries, but also between regions/ provinces within countries (see Figure 3). This dataset of around 120,000 wines broken down by their prices, ratings, countries and provinces should be sufficient to answer my research question.

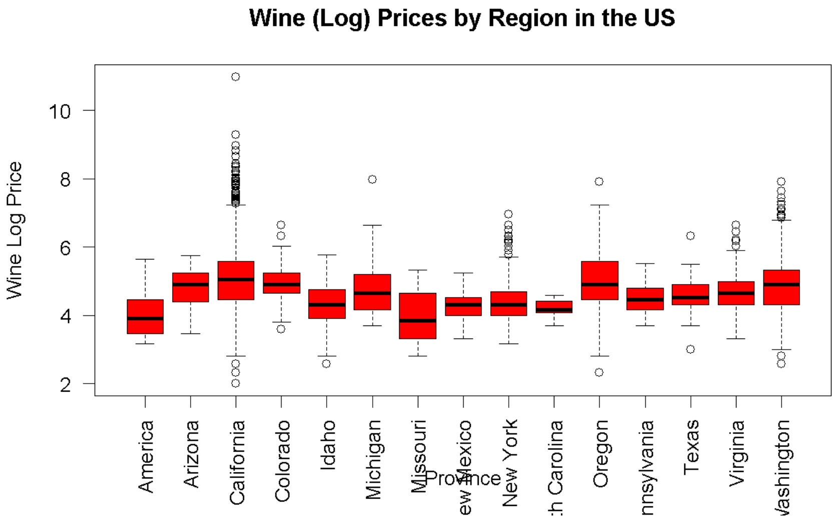

***Figure 3**. Wine (Log) Prices by US Region Boxplot*

Methodology

Bayesian inference is an alternative to the standard frequentist inference which utilises prior information relevant to an event which reflects our beliefs about the variables before seeing the data. Bayesian inference's main features include: prior information (all experiments have context which are expressed by their prior distribution), subjective probability (results depend on subjective knowledge and choice of prior distribution), and uncertainty (a prior distribution is assigned to reflect the uncertainty, not because the truth is random). The standard steps of conducting Bayesian Inference are: 1. Define the likelihood (same as frequentist), 2. Define / calculate the prior, 3. Calculate the posterior (which is proportional to the likelihood multiplied by the prior, and 4. Conduct inference from the posterior. It can be applied for both linear regression and classification problems, and has many real-world applications such as medicine, finance, and scientific research. In many real life applications, Bayesian inference was difficult to apply due to the complex computational aspect of it but the development of Markov Chain Monte Carlo (MCMC) algorithms has made the computation much faster and easier, and can be used for non-standard and significantly complex applications where the classical frequentist approach is infeasible.

MCMC is a way to repeatedly sample from the posterior distribution in order to estimate its parameters which can then be used for inference. MCMC is based on Markov chains which is where for a sequence of dependent, random variables $\{x_t\}_{t=1,\ldots,N}$, the following is true: $\pi(x_{t+1} \mid x_1, \ldots, x_t) = P(x_{t+1} \mid x_t)$ where $\pi(.)$ and $P(.)$ represent probability distribution functions. This means that in a Markov chain, the probability of the next variable x_{t+1} given all the previous variables $x_1, \ldots, x_t$ is the same as the probability of the next variable x_{t+1} given the most recent variable x_t.

Another important aspect of a Markov chain for MCMC is if it has ‘stationary’ ‘transition probabilities’. A transition probability $T_t(x_t, x_{t+k})$ is defined as follows, $T_t(x_t, x_{t+k}) = P(x_{t+k} \mid x_t)$. When a transition probability is independent of time, the Markov chain is called ‘homogeneous’. For a homogeneous Markov chain with transition probabilities $T(x', x)$, the distribution $\pi^*(z)$ is stationary if $\pi(x) = \sum T(x', x)\pi^*(x')$ *over all x’*. If a Markov chain is ‘reversible’ ($\pi(x_t)P(x_{t+1} \mid x_t) = \pi(x_{t+1})P(x_t \mid x_{t+1})$), then its summation (or integration for continuous Markov chains) over all x_t satisfies the stationarity condition. To implement MCMC, you must construct Markov chains with their posterior as stationary, which is still possible if you only know the likelihood and the prior, and then you can use the Markov chains to sample from their stationary distribution. There are three main MCMC algorithms to choose from: Metropolis-Hastings (MH), Gibbs sampler, and Hamiltonian MCMC (HMC). MH will sample parameters from the posterior and then update the parameters with a probability based on the ratio of the proposed new distribution to the posterior distribution whereas Gibbs will sample each of the parameters sequentially, conditional on the other parameters being fixed. However, HMC uses gradient information to update parameters, it can be applied to any model (unlike MH or Gibbs), and it can be implemented with relative ease using Python or R packages such as `rstan`. I have chosen to use a variant of HMC called NUTS (No-U-Turn Sampler) which is slightly different to standard HMC as it automatically selects one of the parameters of HMC to avoid unnecessary work (Stan 2025), run on the `stan_glmer()` function in the `rstan` R library. I will run `stan_glmer()` on the formula, `log_price ~ rating_c + (1 | country / province)`, with the following prior information, ‘`prior = normal(0,1)`’, and ‘`prior_intercept = normal(mean(wine$log_price),1)`’.

I will compare the results of HMC against non-Bayesian methods, such as lasso and ridge regression on the formula, *`log_price`* ~ *`rating_c + country + province,`* which both work by minimising the residual sum of squares (RSS) plus a penalty term proportional to the sum of the square of all model coefficients for ridge (*L2* regularisation) or the sum of the absolute value of all model coefficients for lasso (*L1* regularisation), using the `MASS` R library. The main difference between them is that for ridge, the coefficients can become very small but never zero, whereas for lasso, they can be set to zero. Moreover, the RSS is the sum of the squared value of the residuals, which are the difference between the actual values and the predicted values. In order to compare these models, I will holdout a test set from the training data and use the test mean-squared error (MSE) and error rates to compare how well they have managed to model the data, as well as evaluate parameter estimates and Bayesian credible intervals (CrI). I will also assess how interpretable each of the types of models are based on their different possible summary statistics as this is the other important aspect of determining the quality of the models for my research aim.

Results

After completing the setup, methods and analyses detailed previously, I have found that lasso and ridge methods of frequentist style linear regression performed well. Lasso only had non-zero (significant) coefficient values for 65 of the 257 possible country and province variable options. This resulted in a Ridge test MSE of 0.463 and a Lasso test MSE of 0.469 to 3 significant figures, with a lambda (selected by grid search) of 0.01 and a test set holdout of 20%.

Variable Name	**Mean**	**Standard Deviation**	**5%**	**50%**	**95%**
(Intercept)	4.57	0.06	4.47	4.57	4.66
rating_c	0.17	0.00	0.17	0.17	0.17
Alsace, France	-0.10	0.09	-0.24	-0.10	0.04
Burgundy, France	0.73	0.09	0.59	0.73	0.86
Loire Valley, France	-0.15	0.09	-0.29	-0.15	-0.01
France	0.08	0.10	-0.08	0.08	0.25
Italy	0.22	0.10	0.05	0.21	0.39
US	0.34	0.09	0.18	0.34	0.49
Australia	0.04	0.12	-0.16	0.05	0.24
England	0.22	0.21	-0.11	0.21	0.58
Canada	0.17	0.17	-0.11	0.17	0.46

Table 3. *Example of some summary statistics of the final model. Where a country is given, the variable name is b[(Intercept) country:country_name], and where a country and province is given the variable name is b[(Intercept) province:country:province_name:country_name]. Also, 5%, 50%, and 95% are the credible intervals for each variable.*

For the Bayesian methods, you can find examples of some of our credible intervals in Table 3. These credible intervals do somewhat demonstrate why a hierarchical Bayesian model is useful here as the distributions of the different countries and their regions can vary greatly. For example, in this pricing model, the mean intercept value for the US is over four times greater than that of France, the standard deviation of intercept values for England is roughly double that of Italy, and within France, Burgundy and the Loire Valley's price intercepts have the same standard deviation but Burgundy's mean is 0.88 (substantially) higher.

Moreover, HMC produced the following plots using 4 chains and 2000 (both chosen heuristically) iterations (Figure 4):

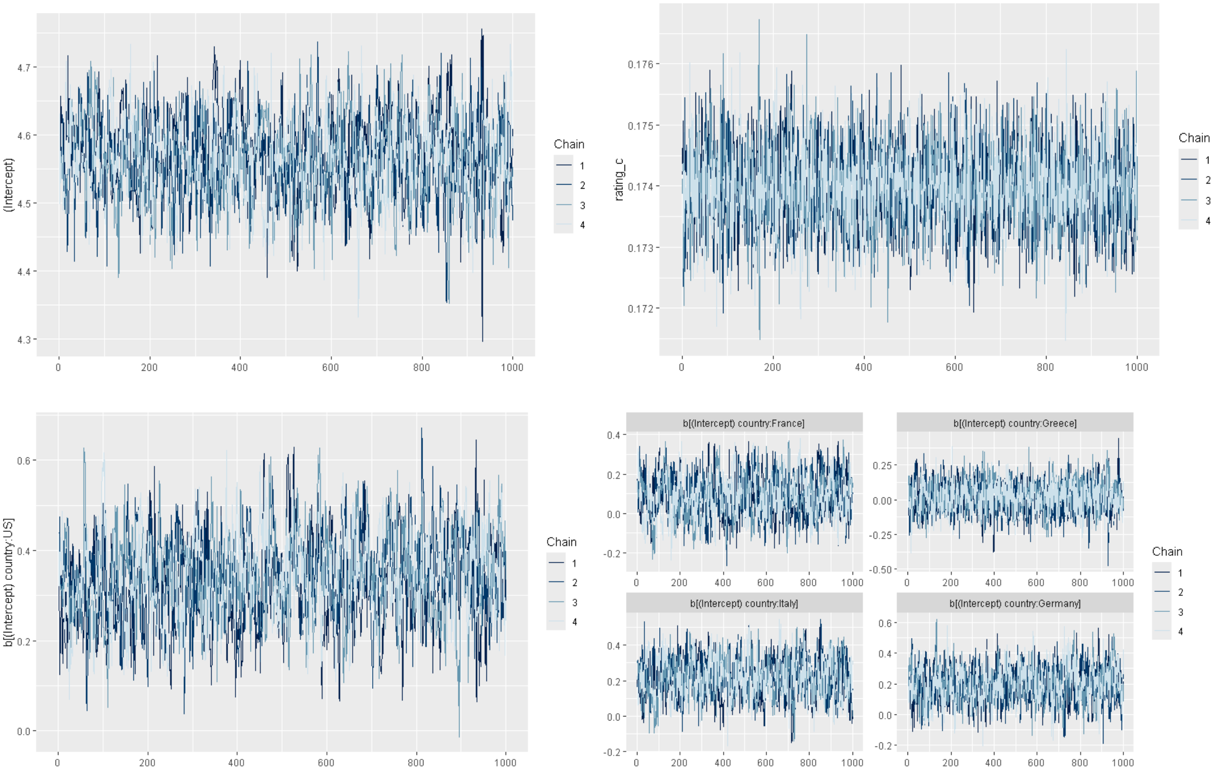

Figure 4. *Trace plots for the intercept, rating_c, US, France, Greece, Italy, and Germany.*

These plots all seem to converge with no upward or downward trends or clear visible divergences, and all chains roughly within the same region. Furthermore, all R-hat values (a measure of Markov chain convergence) were around 1.0000 with the minimum being 0.9993 and the maximum being 1.0210 (to 4 decimal places due to small magnitude of variance) and so the chains have converged. Also, the posterior predictive check (not shown here but available in the code) showed the model predictions appearing to be similar to the shape and size of the actual data and they seem to follow the same distribution roughly. However, the use of the discrete variable `rating_c`, has made the model slightly less smooth especially around the mean(s) and so if I were to do this model again I would find a way to better include this ordinal, discrete variable.

Discussion

One limitation of the data set I have used is that the amount of data towards the US is significantly larger than that of France, Spain or Italy, despite those countries producing more wines than the US and so the data and therefore the model may have some bias towards the US and so have worse generalisation. Also, one limitation of HMC is its sensitivity to the choice of prior(s). For my research, I chose relatively uninformative priors as I had less prior knowledge of how the priors might be distributed.

One of the advantages of using a Bayesian method is the ability to provide full posterior distributions over the parameters, which allows for the calculation of credible intervals and so better uncertainty modelling compared to Lasso and Ridge which have point estimates and asymptotic confidence intervals. Furthermore, being able to include prior information and so extra knowledge about the data is useful for data which is noisy, and isn't used in frequentist methods. The flexibility of HMC enables it to handle complex models, like the hierarchical one I have used, whereas most frequentist methods would likely just assume a simpler, linear model. Also, the ability to evaluate the model, using posterior predictive checks (`pp_check()` in R) is exclusive to Bayesian methods like HMC. However, HMC is very computationally expensive (it took 6 hours to run on my reasonably modern computer, see Appendix for more details), especially for complex models and large datasets whereas Ridge and Lasso scale better for larger datasets and so can model them much faster. Moreover, while the posterior is very detailed, it is also harder to explain concisely, unlike the more easy to interpret and explain, simple point estimates for Lasso and Ridge regressions. Also, Lasso has the ability to reduce the number of necessary variables by reducing some coefficients to zero.

For further research, I would find confidence intervals for the frequentist methods and compare them to the Bayesian credible intervals, as well as find a more balanced dataset for a variety of countries and their regions.

Conclusion

The main findings from our results is that the price of a wine and how the prices of wines from a specific region or country are distributed are well modelled using Bayesian methods like HMC as its ability to model complex, hierarchical models and ability to utilise prior knowledge allows it to create interesting and accurate descriptions of how wine pricing distributions vary between countries and their provinces/ regions. Moreover, one of the main takeaways from this article is that while both Bayesian and frequentist methods model the data well, the inference for Bayesian models with its prior knowledge seems to give a richer understanding of the data which while harder to interpret concisely is much more in-depth. For further research, I could look into ways to model ratings based on price as ratings are discrete, not continuous, and they are ordinal (a wine with a score of 90 is better than a wine with a score of 80) and so would not be able to be modelled by a gaussian, binomial, or poisson distribution effectively.

References

(Stan 2025)

Stan Development Team. (2025). *Stan reference manual*. https://mc-stan.org/docs/reference-manual/

Appendix

- Values were given to 2 decimal places where necessary, unless otherwise stated.
- Ran on a computer which has a Ryzen 7 5800X and 32GB of DDR4 ram.
- Code attached in file: ‘`st308_proj_script.Rmd`’