Wine Review

Matthew Beebe, Tianqing Bei, Anneliese Johnson, Srihan Kakarlapudi

Stat 365 - Final Report

December 10, 2024

Wine Review Data:

Inspired by *Somm,* a documentary on master sommeliers, this Kaggle dataset has 130k wine reviews. The data contains 10 attributes:

1. #, for the index.
2. country, the country that the wine is from.
3. description, which contains a brief description of the wine (119,955 unique values!).
4. designation, the vineyard within the winery where the grapes that made the wine is from.
5. points, the number of points WineEnthusiast rated the wine on a scale of 1-100.
6. price, the cost for a bottle of wine.
7. province, the province or state that the wine is from.
8. region_1, the wine growing area in a province or state
9. region_2, a more specific region, if specified.
10. taster_name, the name of the taster, if provided.
11. taster_twitter_handle, the twitter handle of the taster, if provided.
12. title, the title of the wine.
13. variety, the variety of the wine.
14. winery, the winery that the wine came from.

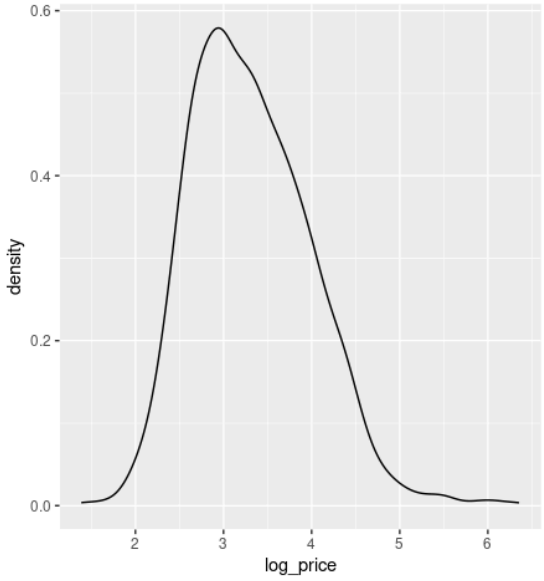

Figure 1: Density plot of the log_price.
$\mu = 3.313282 \qquad \sigma = 0.6791097$

Additionally, we added a variable "log_price" which took the log value of the wine prices. This helped to better understand the true distribution of prices by standardizing the values. However for our regression model, we tried to predict the actual price not the log price; the log price just helped to better understand the distribution.

To understand the data properly, the data was filtered to only use data from countries with over 400 reviews. This resulted in wine reviews from Argentina, Australia, Austria, Chile, France, Germany, Italy, New Zealand, Portugal, South Africa, Spain, and the US.

The wines were distributed across the countries as follows:

1. Argentina: 0.032921811
2. Australia: 0.015432099
3. Austria: 0.021604938
4. Chile: 0.046296296
5. France: 0.021604938
6. Germany: 0.021604938
7. Italy: 0.136831276
8. New Zealand: 0.136831276
9. Portugal: 0.045267490
10. South Africa: 0.009259259
11. Spain: 0.48353909
12. US: 0.4444444444

‘

Country	Mean Points	Mean Log Price	n
Argentina	86.1	2.98	32
Australia	88.7	3.19	15
Austria	89.6	3.13	21
Chile	86.5	2.76	45
France	89.3	3.46	164
Germany	90.5	3.44	21
Italy	88.8	3.40	133
New Zealand	88.4	3.13	9
Portugal	88.3	2.93	44
South Africa	88.2	3.31	9
Spain	86.9	2.96	47
US	88.4	3.40	432

Figure 2: Summary of the wine data by country with specific focus on mean points and mean log price

Due to the computing constraints of posit cloud, we decided to use a n = 1000 sample of the filtered data so our MCMC simulations could run. We used seed = 36501 to preserve the integrity of the saved data.

Bayesian Estimation:

This part summarizes the Bayesian analysis conducted on the points variable from a wine review dataset. The objective was to estimate the posterior distributions for the mean (mu) and standard deviation (sigma) of points, leveraging a normal likelihood with unknown parameters. The analysis was performed using the STAN framework, which enables efficient Markov Chain Monte Carlo (MCMC) sampling.

Data Preprocessing and Exploration

The points variable represents numerical ratings assigned to wines. Before analysis, the data underwent preprocessing:

Missing Data: Observations with missing points were excluded to ensure reliable inferences.

Descriptive Statistics:

Mean: ~ 88

Standard Deviation: ~ 3

Range: Approximately 80 to 100

The distribution of points appeared unimodal and approximately symmetric, justifying the use of a normal likelihood model.

MCMC Sampling:

The posterior distribution was sampled using the No-U-Turn Sampler (NUTS) algorithm in STAN:

Number of Chains: 4

Iterations per Chain: 6000 (3000 warm-up, 3000 sampling)

Total Samples: 12,000

The Bayesian model applied to the dataset is as follows:

Likelihood**:** Observed data yi ~ N(μ, σ^2), where yi is the point for the i-th wine.

Priors were chosen to reflect moderate prior knowledge:

$\mu \sim N(85, 5^2)$: Centered around a plausible mean score (85), with a wide spread to account for uncertainty.

σ ~ Half-Cauchy(0, 2.5): A weakly informative prior to constrain sigma to reasonable values while avoiding overly strong assumptions.

Posterior**:** Posterior distributions for μ and σ were inferred via MCMC sampling.

Inference and Results

The MCMC sampling generated 12,000 posterior samples from 4 chains, ensuring convergence and sufficient sampling for reliable estimates.

Parameter	Mean	Standard Deviation	95% Credible Interval (CI)	Effective Sample Size (ESS)	R-hat
mu	88.464	0.106	(88.254, 88.674)	10,943	1.00

sigma	3.248	0.073	(3.110, 3.393)	11,255	1.00
lp__	-1629.362	NA	NA	5,470	1.00
Figure 3: Posterior summaries for the parameters					

The posterior mean of μ suggests that the average wine rating is approximately 88.464. The tight credible interval reflects high confidence in this estimate. The posterior mean of σ indicates moderate variability in the ratings, with a standard deviation of approximately 3.248. This indicates stable and precise estimates of the variability in reaction times.

The mean log-posterior probability is approximately -1629.362. This value serves as an indicator of the model's goodness of fit to the data (higher is better, but this is primarily used for diagnostic purposes).

The posterior distributions for the reaction times (μ_i) were successfully estimated for all individuals.

Summary statistics for μ_i:

- Mean Reaction Time: Ranged across individuals with clear variation.
- Uncertainty (Credible Intervals): Larger for individuals with fewer observations, consistent with the no pooling assumption.

Individual-Level Variation:

- Each country's posterior mean for μ_i was distinct, highlighting significant variability in reaction times.
- Subjects with more data had tighter posterior credible intervals, reflecting higher certainty in the estimates.

No Pooling Effects:

- The lack of shared information led to high flexibility in capturing individual-specific effects.
- However, individuals with sparse data exhibited wider uncertainty, underscoring the trade-off in the no pooling approach.

Interpretation

The Bayesian no pooling approach provided individual-specific estimates for reaction times, allowing for highly personalized insights. This is particularly useful in scenarios where the assumption of homogeneity (as in complete pooling) is unrealistic. However, the model's flexibility comes with limitations:

- High Variance for Sparse Data: Subjects with fewer observations had less precise estimates.
- No Borrowing Strength: The model does not leverage shared information across individuals, which could be advantageous in data-limited settings.

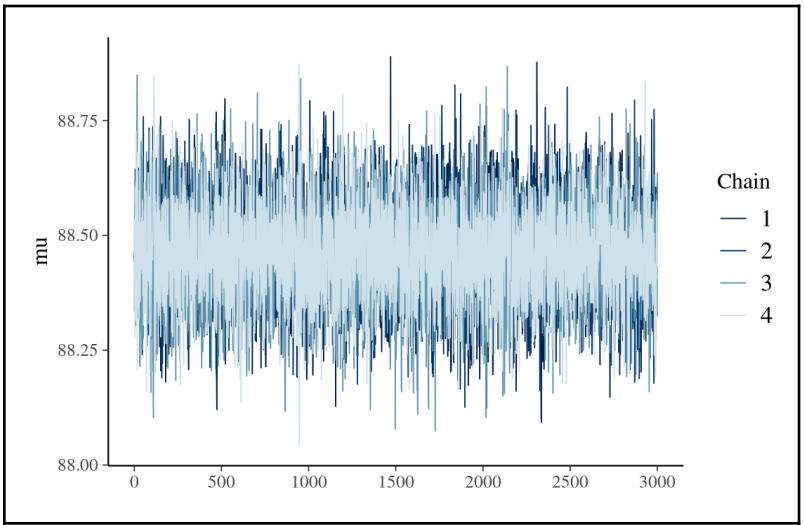

Figure 4: Trace Plot of the MCMC

The trace plot for the parameter μ displays the MCMC sampling progression across four chains. The chains mix well, with no discernible patterns or trends over iterations. This indicates that the chains have likely reached the stationary distribution, a critical requirement for valid inference. All chains overlap significantly, showing that different initial values converged to the same posterior distribution. This provides confidence that the MCMC process did not get stuck in local modes. The samples oscillate randomly around a consistent range, further supporting good convergence.

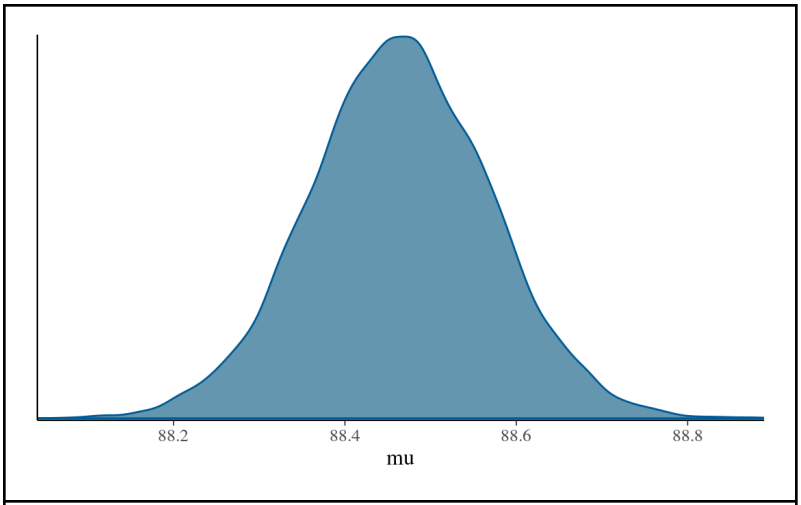

Figure 5: Posterior Density Plot

The density plot provides insight into the posterior distribution of μ. The density is smooth and unimodal, indicating that the posterior distribution is well-sampled and does not have multiple modes. The posterior distribution has a relatively tight spread, which reflects a high degree of certainty in the posterior estimate of μ. The peak of the density corresponds to the posterior mean of approximately 88.46, which is the most probable value of the mean reaction time.

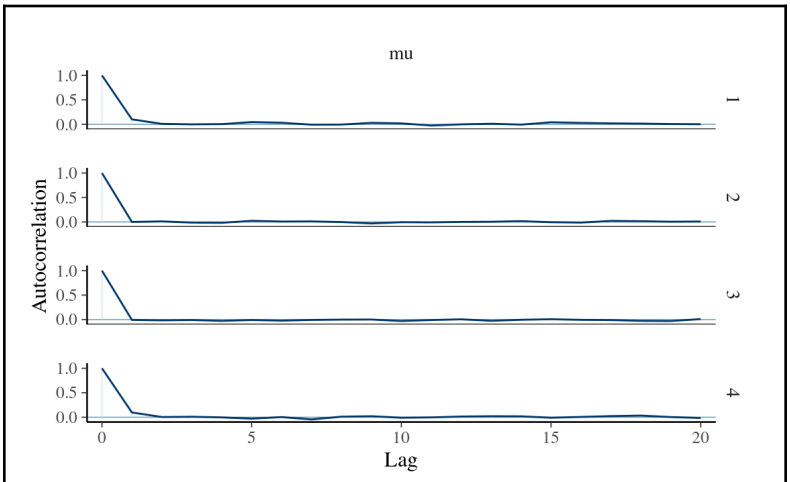

Figure 6: Autocorrelation plot of the MCMCs

The autocorrelation plot measures the dependence between samples in the MCMC chain as a function of lag. The autocorrelation decreases rapidly and approaches zero within a few lags. This indicates that consecutive samples are mostly independent, which is crucial for efficient sampling. Minimal

autocorrelation at higher lags suggests that the chains are exploring the posterior space effectively without getting stuck in regions of high correlation.

Bayesian Regression:

For our Bayesian regression model, we used the generalized linear models via Stan, taking a hierarchical approach. The data was grouped by the country it was from, as we noticed the trend that prices had depended on the country it was from (e.g. all of the most expensive bottles were from France). We hypothesized that there was likely a relationship between the rating of the wine (predictor) and the price of the wine (response). For the generalized linear model, we used prior knowledge of the global mean price of wine ($\mu = 35$), and the likely interval for price (between 22 and 48). This provided us a prior for the intercept that was $N(35,\ 6.5)$. For our slope, we used a prior belief that wine increases around $3 per point. We used a weakly informative prior for our variance. The model ran four MCMC chains with a total of 8,000 interactions each (4,000 warm-up, 4,000 sampling). The code took significant computational effort to function, which showed through the elapsed time for each chain ($\mu_{\text{Elapsed Time}} =$ 594.263 seconds). The model diagnostics that we used included effective sample size with *neff_ratio,* and r-hat. Since the wine model was grouped by country, the *neff_ratio* function provided us the ratio of the effective sample size to the total number of posterior samples for the intercept and the slope of all 12 countries.

Country	**Intercept Neff Ratio**	**Slope Neff Ratio**
Argentina	0.8719375	0.8772500
Australia	0.7736875	0.7765625
Austria	0.8469375	0.8705000
Chile	0.9246875	0.9261250
France	0.4851250	0.4837500
Germany	0.8975000	0.9024375
Italy	0.5588125	0.5568750

New Zealand	1.1824375	1.1868125
Portugal	0.8435625	0.8409375
South Africa	1.2060000	1.2114375
Spain	0.7600625	0.7595000
US	0.4158125	0.4146875
Figure 7: Chart of the N-effective ratio by country of the wine review model.		

A value close to 1 suggests good mixing and minimal autocorrelation, indicating high efficiency. We found that the residual standard deviation had a ratio of 1.3263125 which represents the average distance of the observed values. Shown by ***Figure 7***, some of the chains performed with high efficiency, such as Chile, Portugal, and Spain, while other chains performed sub-par, like Argentina. This may call for a shift in our priors to get them to appropriately aid in convergence.

Country	**Intercept R-hat**	**Slope R-hat**
Argentina	0.9999006	0.9998939
Australia	0.9999502	0.9999511
Austria	1.0002796	1.0002703
Chile	0.9998688	0.9998613
France	0.9998825	0.9998788
Germany	0.9999544	0.9999518
Italy	0.9999865	0.9999866
New Zealand	0.9998559	0.9998538
Portugal	0.9998312	0.9998305
South Africa	0.9998896	0.9998935
Spain	0.9998800	0.9998775
US	1.0001803	1.0001654
Figure 8: Chart of the R-hat by country of the wine review model.		

The R-hat value assesses whether the MCMC chains from our model have converged. It does this by measuring the ratio of the variance between chains to the variance within the chains. Values close to 1 indicate good convergence. All of the R-hat values from our data model are very close to 1, which suggests that the estimates are reliable.

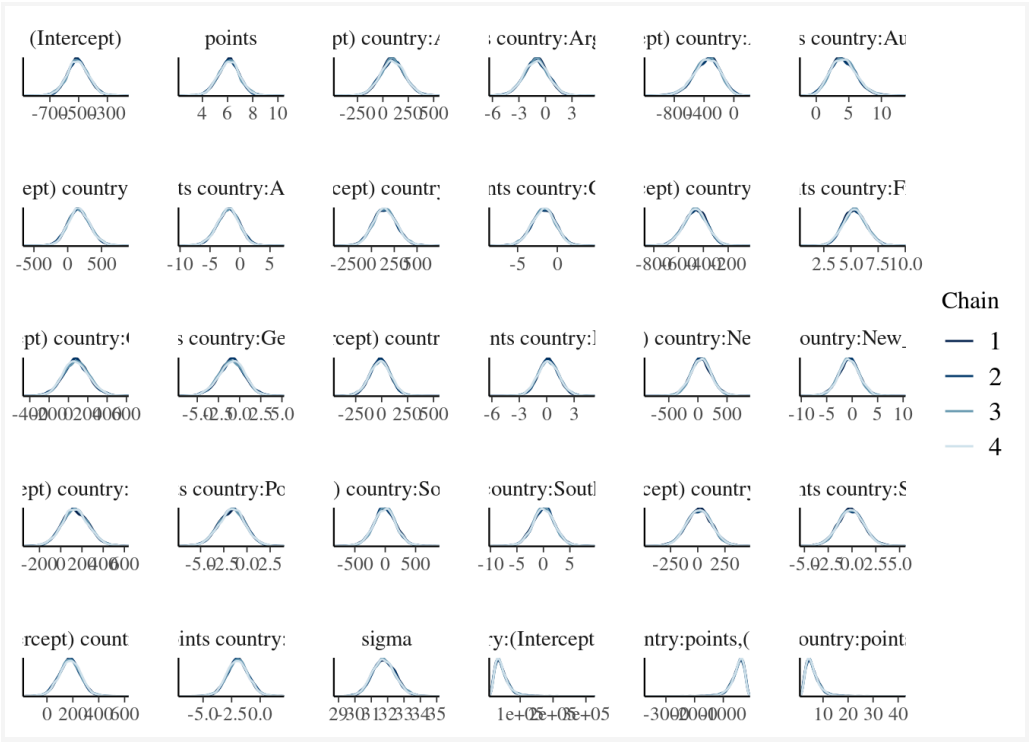

Figure 9: Monte Carlo Markov Chains in a density overlay.

Figure 9 shows a visualization of the MCMC simulation diagnostics, providing an overlay plot of the multiple MCMC chains. This plot shows consistency in the density shapes across the country groups that indicate that the chains have converged to the same prior distribution.

Figure 10: MCMC Trace Plots

The trace plots help as a diagnostic tool for the MCMC simulations. These trace lines show that the MCMCs have good convergence through their good mixing. There seems to be no unusual patterns or trends, which indicates good sampling and testing.

To visualize the data at hand, we created scatterplots with simple regression lines that looked at the correlation between price and points by country. This visualization showed us the variance of slope and intercept by country rather than globally. In this figure, we see that countries like New Zealand and South Africa have quite steep regression lines (high B_1 values) but pretty limited sample size. We can also see that countries like the US and France have lower B_1 values, but much more samples.

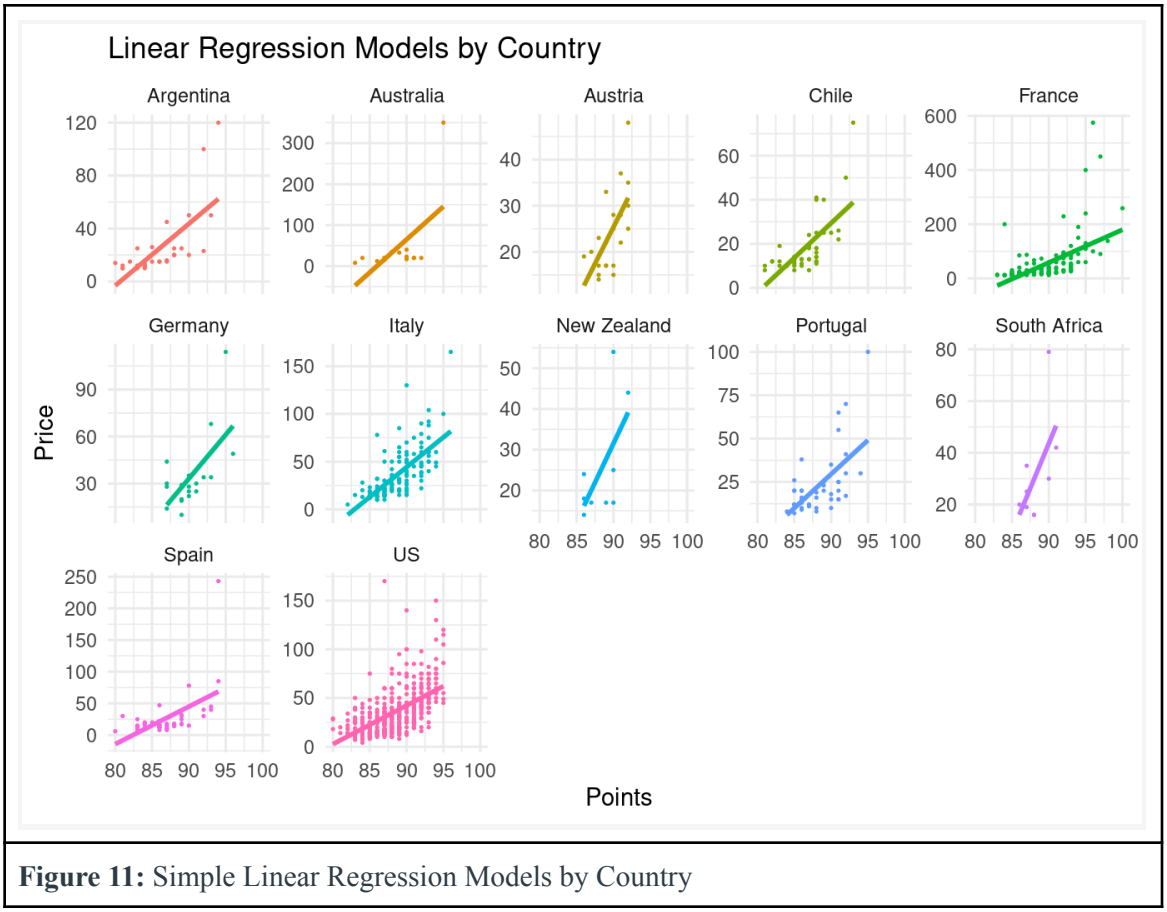

Figure 11: Simple Linear Regression Models by Country

For the fixed effects of our global (countries all combined) regression model, we found that the intercept would be -$508 (CI: [-632, -378]) when points are 0, however when using a centered point value, the mean point rating: 88, we get $29.68, which is a reasonable estimate for the average wine. The slope for points is 6.11 (CI: [4.63, 7.53]) which confirms a strong positive association between price and points. The credible intervals suggest that the effect of points on price is robust.

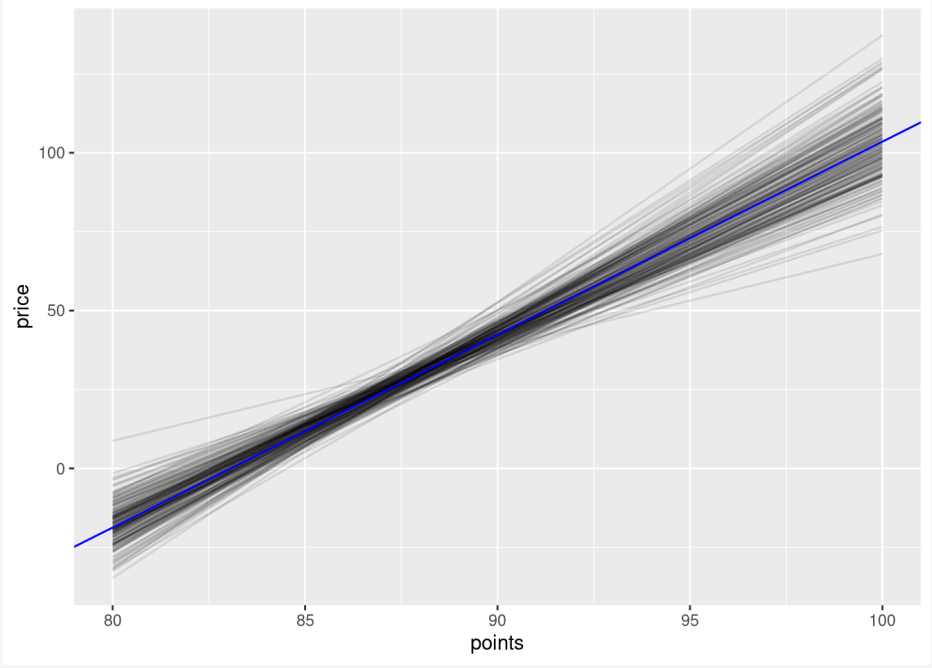

Figure 12: Plot of the linear regression line and fitted draws (n = 200) for all of the wine data, not country specific.

This plot visualizes the uncertainty in the posterior predictive distribution of our model. It works by generating 200 posterior draws of the values from the model, representing a plausible set of parameter values for the model, and overlays the line that is the mean estimate of the model. As we can see, we have a higher confidence on price estimations around the 85 to 90 points range, where the spread of lines is more tightly packed. There is more uncertainty towards the higher point values and the lower point values, however.

In order to get country-specific intercepts and slopes, we created MCMC chains by country, and then generated posterior predictive lines to visualize the data.

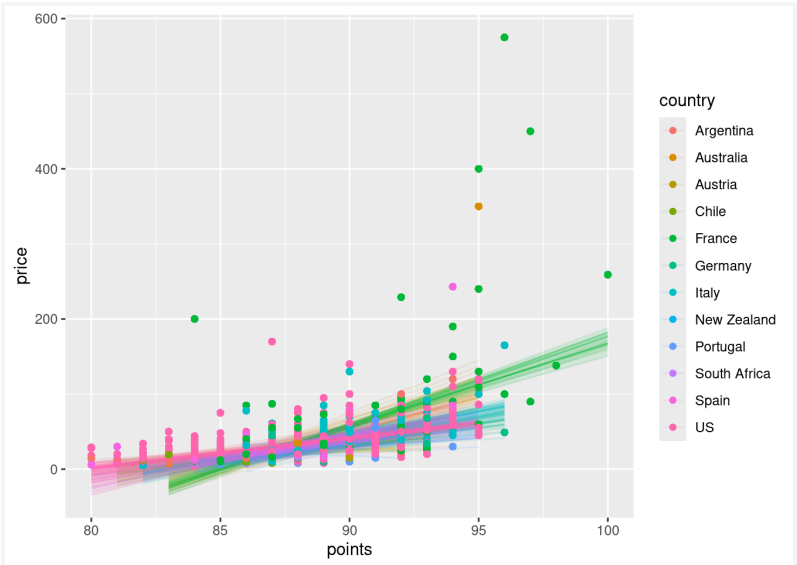

Figure 13: 50 plausible models for each country.

Each of these lines represents a regression model sampled from the posterior distribution for a specific country, and the spread indicates the uncertainty in predictions. This fit across the countries shows how the relationships vary pretty significantly both in point value and price.

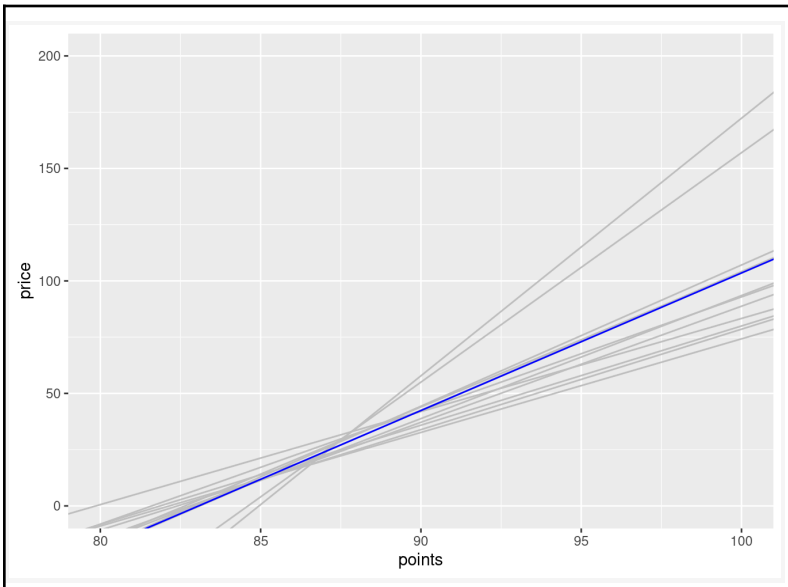

Figure 14: Plot of the country specific posterior median models (grey) with the global median model (blue).

To gather more insights from the plot, we plotted the country-specific posterior median models next to the global median model. This allows us to examine the differences in intercepts and slopes to show how the relationship between *points* and *price* vary by country.

Country	Intercept	Slope
Argentina	-412	5.05
Australia	-863	10.2
Austria	-342	4.16
Chile	-361	4.41
France	-973	11.4
Germany	-427	5.16
Italy	-521	6.28
New Zealand	-453	5.47
Portugal	-368	4.46
South Africa	-512	6.16
Spain	-494	5.97
US	-330	4.14
Figure 15: Summary of the intercepts and slopes for individual countries.		

This summary shows the intercept and slope values for the regression lines of each country. Theoretically, this implies that a wine given 0 points would cost negative amounts of money, which isn't realistic, and shows how the certainty of our model is much less at different values. All of the samples ranged from point value 80 to 100.

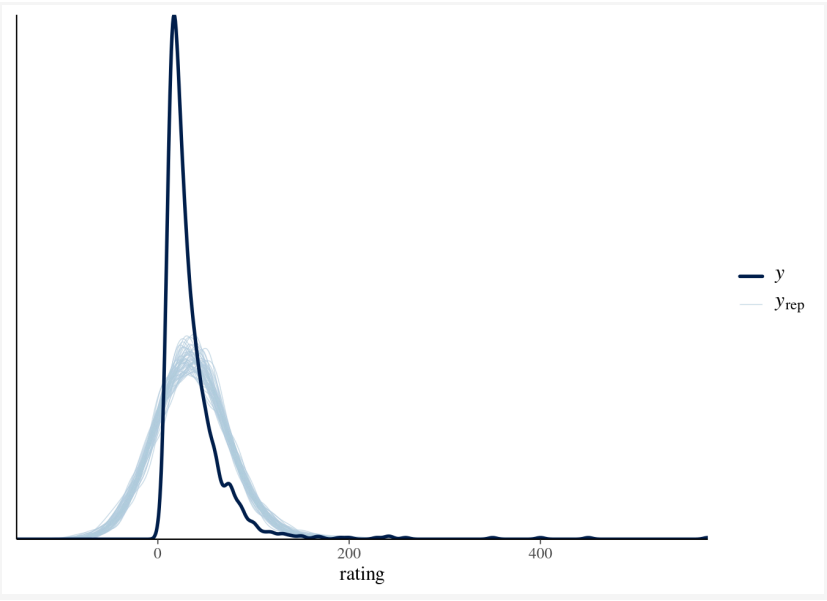

Figure 16: Posterior predictive check with the observed data.

In order to check our prediction, the *pp_check* function. This compared our model predictions with the actual distribution of the observed points. The plot shows that our predictions have a higher variance. This indicates the model has some uncertainty in its ability to replicate the observed data precisely. The close mean, on the other hand, indicates that the model is reasonably accurate in capturing the central tendency of the data. We found that the mean absolute error (MAE) was 11.75247, which indicates how far, on average, the predictions were from observed prices. While this seems high, our scaled MAE is closer to 0, at 0.3665858. This suggests smaller errors relative to the data range. We also found that 78.7% of the observed values fall within a 50% credible interval of our model's predictions. (And 97.9% within the 95% credible interval!).

In order to cross-validate our predictions, we used the *loo* function. This function performs Leave-One-Out (LOO) cross-validation which evaluates the model's predictive performance by systematically leaving out one observation at a time and calculating how well the model predicts the left out data. The output of the LOO cross-validation provides estimates of model fit and complexity. The cross-validation estimated an Expected Log Predictive Density (ELPD) of -4771.9, which suggests moderate predictive performance. The Effective Number of Parameters (*p-loo)* was 62.2. This value is used to reflect the model complexity. A smaller value relative to the total number of observations indicates less risk of overfitting. Our value of 62.2 indicates moderate complexity. The final value that we received from the cross validation was the LOO-Information Criterion, which we received a score of

9543.7. Lower values indicate better predictive performances, which indicates that our model could be better at prediction.

In the loo function, there were 3 observations that were noted to have a large influence on the model and were noted to affect the validity of the LOO estimates. Going forwards, re-running the loo with a k threshold of 0.7 could refit the model and obtain more accurate estimates.

Discussion:

Through the analysis of our estimation and regression models, we have identified key insights in the relationship between price, points, and country of origin. Through our estimation, we found a posterior mean rating of 88.46, and standard deviation of 3.25. We confirmed these estimates with MCMC diagnostics and credible intervals.

The regression model that we performed observed a positive relationship between points and price globally and within individual countries. This confirmed our initial hypothesis. The global model estimated a slope of 6.11 (CI: [4.63, 7.53]), however country specific intercepts and slopes varied. The steepest slope was observed to be France at 11.4, while the lowest slope was the US at 4.14. To test our model, we used MCMC trace plots, density overlays, and autocorrelation plots to confirm effective sampling and convergence. We also used LOO cross validation to indicate moderate predictive performance. This could be due to outliers in price.

Our model was limited by the variability in data points by country, as well as outliers which may have skewed results. Another thing to note is that the negative intercept values we arrived at, while they don't seem realistic, a wine score of 0 is also unrealistic. A future application of this work would be to work with transformations to try to fit the data.

References:

Johnson, A. A., Ott, M. Q., Dogucu, M., *Bayes Rules! An Introduction to Applied Bayesian Modeling,* 2021. https://www.bayesrulesbook.com/

Zackthoutt, "Wine Reviews." Kaggle. Updated 2017. https://www.kaggle.com/datasets/zynicide/wine-reviews

Appendix:

Figure 1: Density plot of the log_price.
Figure 2: Summary of the wine data by country with specific focus on mean points and mean log price.
Figure 3: Posterior summaries for the parameters
Figure 4: Trace Plot of the MCMC
Figure 5: Posterior Density Plot
Figure 6: Autocorrelation plot of the MCMCs
Figure 7: Chart of the N-effective ratio by country of the wine review model.
Figure 8: Chart of the R-hat by country of the wine review model.
Figure 9: Monte Carlo Markov Chains in a density overlay.
Figure 10: MCMC Trace Plots
Figure 11: Simple Linear Regression Models by Country
Figure 12: Plot of the linear regression line and fitted draws (n = 200) for all of the wine data, not country specific.
Figure 13: 50 plausible models for each country.
Figure 14: Plot of the country specific posterior median models (grey) with the global median model (blue).
Figure 15: Summary of the intercepts and slopes for individual countries.
Figure 16: Posterior predictive check with the observed data.