Performing Bayesian analysis for prediction of quality score for red and white variants of Vino Verde.

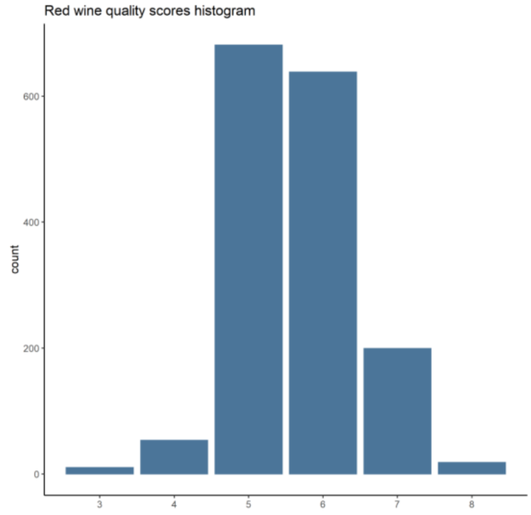
Preparer: Alexey Pankratov

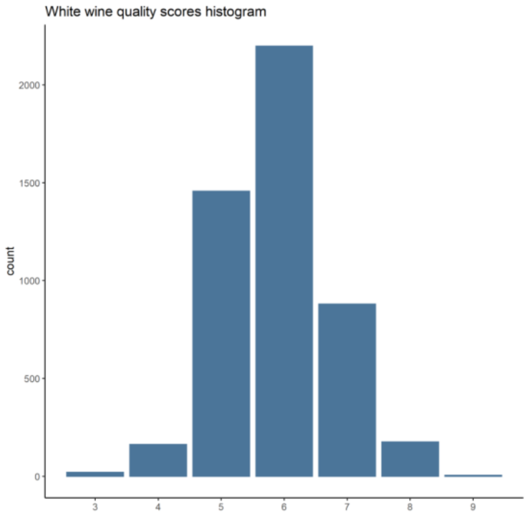
Introduction

The aim of this project is to perform Bayesian analysis for 2 datasets of Vino Verde (red and white variants). Both datasets contain one response variable "quality" that intended to be ranging from 0 (poor quality) to 10 (outstanding quality) but in fact it takes valued from 3 to 8 for red wines and from 3 to 9 for white wines. Both datasets have 11 exploratory variables that will be presented in the plots below. We will first perform the analysis assuming the Normal response model with original explanatory variables form the dataset, then we will transform some of the variables and reperform the analysis. After that, we will perform Bayesian analysis under assumption the response variable 'quality' is distributed in accordance with Multinomial distribution.

Exploratory analysis of the data

We will first have a look at the distribution of quality scores for white and red wines. Red wine dataset has 1599 samples of wine, whereas whites wine dataset has 4898 samples. As it can be seen from histograms below, scores are not evenly distributed.

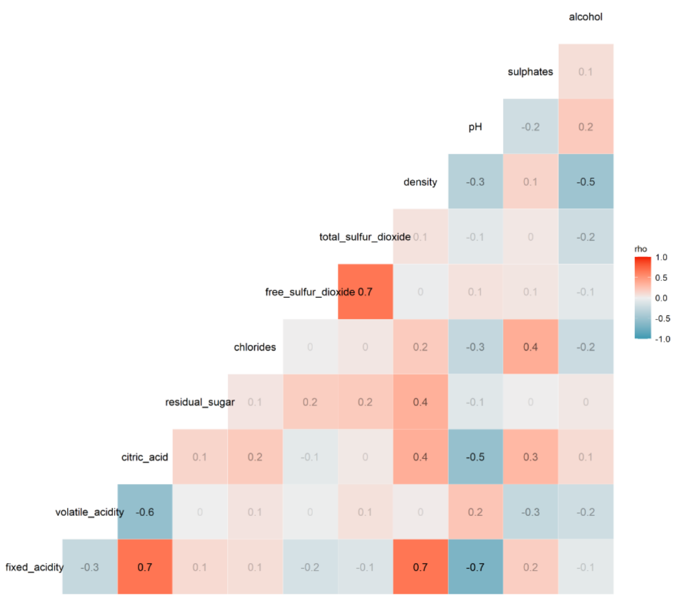

Red wine quality scores histogram


White wine quality scores histogram

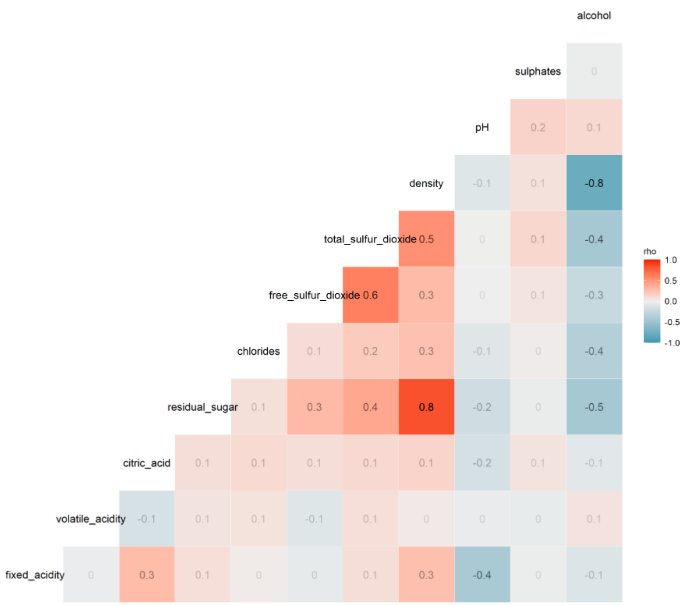
The is a significant number of wines with "average" scores of 5 and 6, much less wines with score 7 and no or almost no wines with very high or very low scores.

And now we will look at the correlation of the 11 explanatory variables in both datasets.

Red wine. Correlation of explanatory variables



White wine. Correlation of explanatory variables



The correlation matrices above show features of each type of wine. For example, it is expected that density will have positive correlation with residual sugar and negative correlation with alcohol as the wine is mostly water solution and water has density of 1.00 g/cm^3, whereas density of sugar is 1.59 g/cm^3 and of alcohol is 0.79 g/cm^3. So, the increase of sugar will increase overall density of wine and increase of alcohol will decrease its density. This correlation is more pronounced for white wines (0.8 correlation with sugar and -0.8 with alcohol) and less for red wines (only 0.4 with sugar and -0.5 with alcohol). Also, it is expected that pH will be strongly correlated with fixed acidity and citric acid as pH is another measure of acidity. Free sulfur dioxide has strong positive correlation with total sulfur dioxide for both wines, which is obvious from the description of the variables (i.e. free is part of the total). At the same time, it is not obvious why fixed acidity has strong positive correlation with density on red wine, which is probably also results from higher density of acids in wine compared to water.

Normal response model – original variables

We will start Bayesian analysis under assumption that quality score data is distributed Normally as follows:

$$y \mid \mu, \sigma \sim Normal(\mu, \sigma^2)$$

Where Y_i is a quality score, standard deviation σ that we assume to be distributed in accordance with Inverse Gamma distribution (explained below) and mean parameter μ_i is a linear function of all 11 explanatory variables (plus intercept):

$$\mu_i = \beta_0 + X\beta$$

Where xi is a vector of 11 explanatory variables (observed) and β is a vector of regression coefficients (unknown). In this project, we assume that β coefficients priors are distributed in accordance with Laplace distribution that can be decomposed into the hierarchical model set our below.

And the full model can be summarized as follows:

$$y \mid \beta, \sigma^2 \sim Normal(\beta_0 + X\beta, \sigma^2)$$

$$\sigma^2 \ \sim Inv - Gamma(a, b)$$

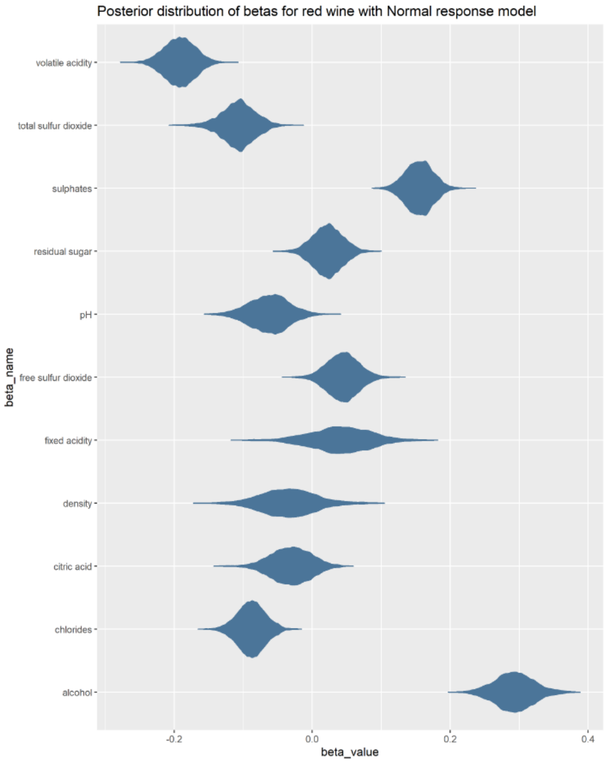
$$\beta_0 \propto 1 \ (constant)$$

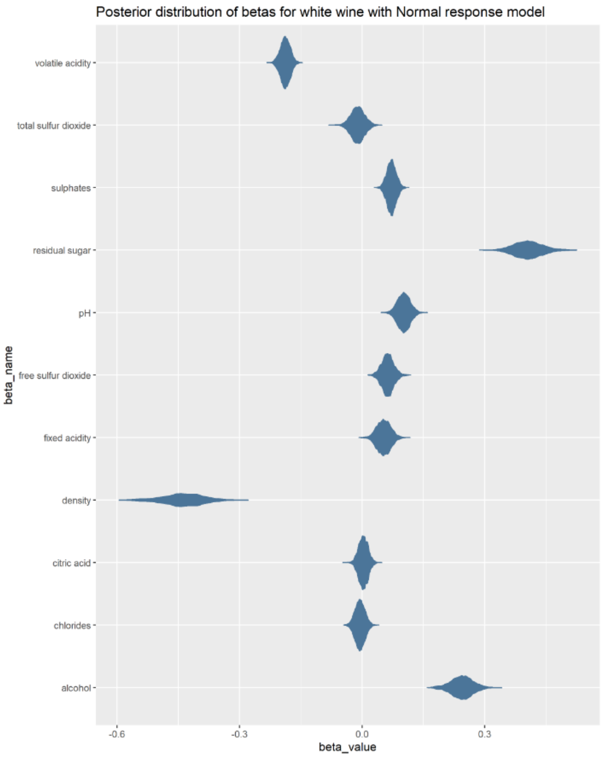
$$\beta \mid \sigma^2, \tau_1^2, \tau_2^2 \dots \tau_{11}^2 \ \sim Normal(0, \sigma^2 D_\tau)$$

$$\tau_i^2 \mid \lambda^2 \ \sim Exponential(0.5\lambda^2)$$

$$\lambda^2 \ \sim Gamma(a_\lambda, b_\lambda)$$

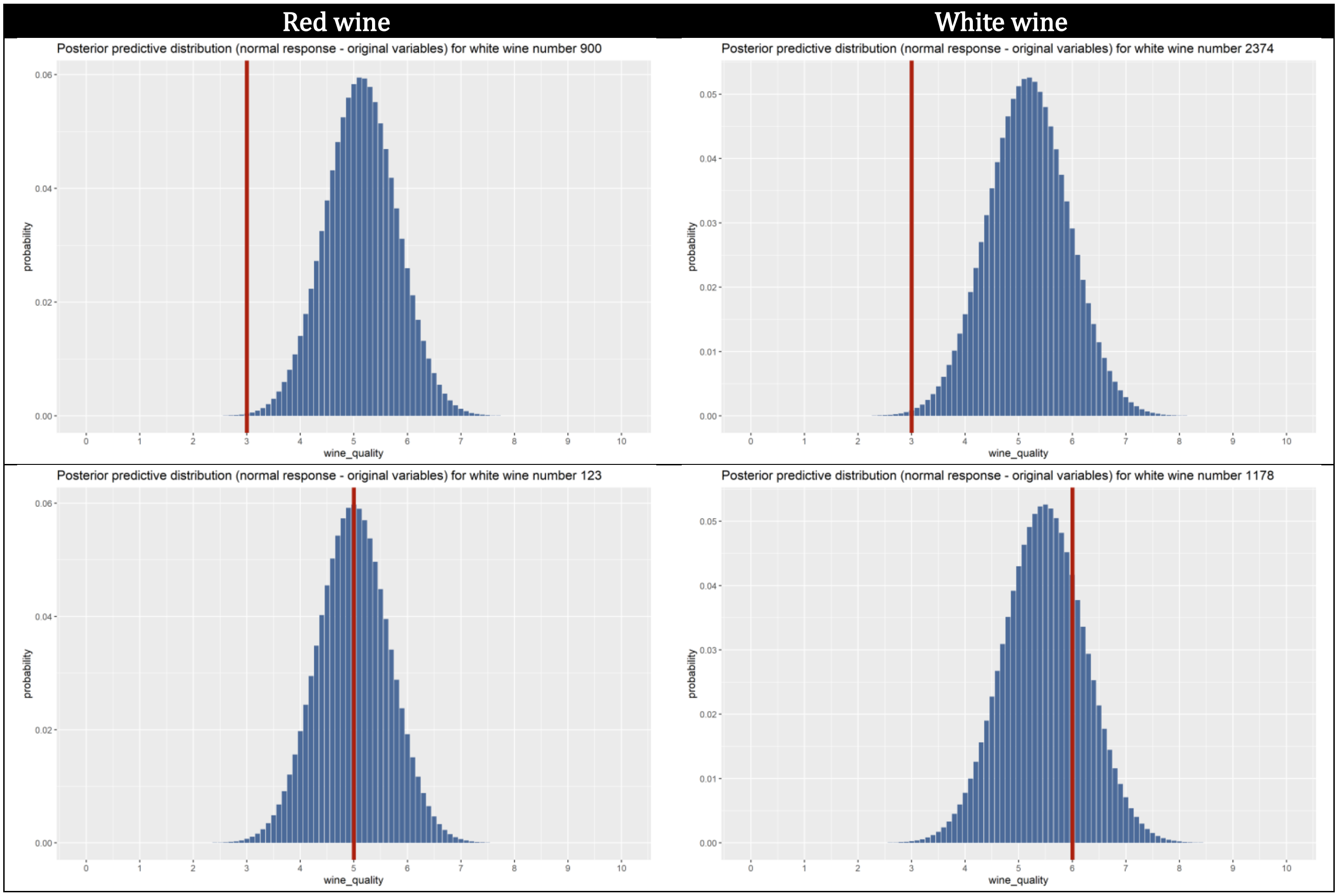
We chose hyperparameters $a = 4 \ and \ b = 60$ for prior σ² and $a_\lambda = 0.025 \ and \ b_\lambda = 0.1$ for prior λ in order to cover reasonable range of likely values of σ² and λ and performed sampling using Metropolis-Hastings sampling method. In total we took 120,000 samples for each wine type and then removed first 50,000 samples leaving 70,000 samples of representing posterior distribution of model parameters β and σ². The violin plot below shows posterior distribution of β for red and white wines. It should be noted that the project I made 3 iterations 120,000 draws each in order to check that posterior distribution converges around the same mode estimate for each iteration. The draws are saved in the folder draws that is submitted together with the other files of the project. The comparison of posterior distribution is presented in the **Appendix 1** there are almost not differences between all three iterations. Below plots represent iteration #1.

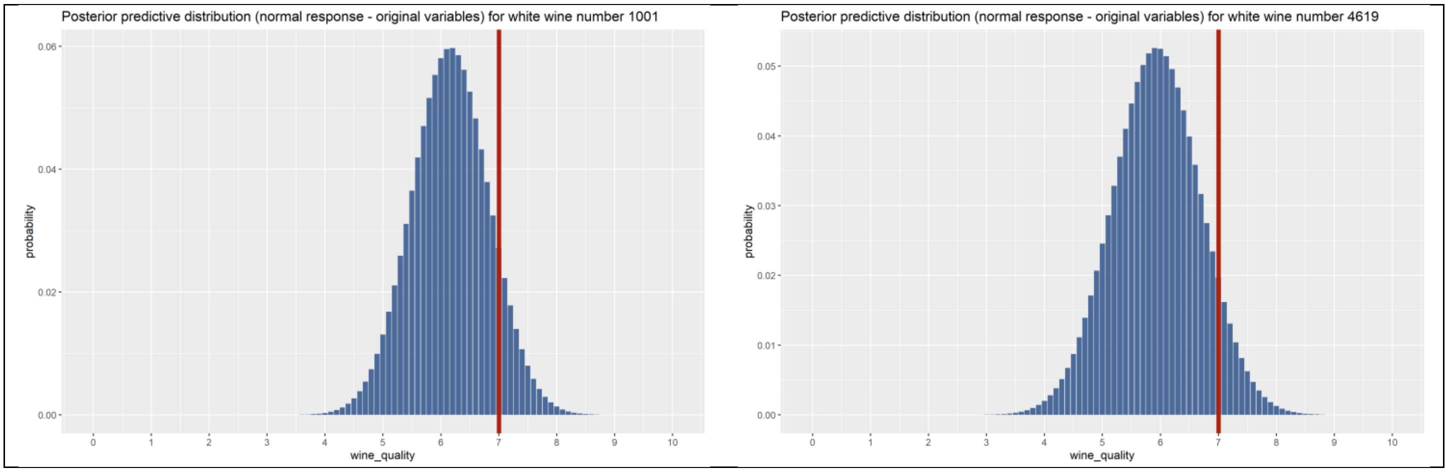




The violin plots of posterior distribution of β parameters show that for both red and white wine alcohol has a significant positive impact on the quality score as mode estimate of β for alcohol is the largest for red wine, approximately +0.29, and approximately +0.25 for white wine. Volatile acidity with mode estimate of β of -0.19 is the most negative effect for red wines and also negative effect for white wines, although not as pronounced. Volatile acidity is associated with vinegary smell so the negative effect on the quality is expected. It also should be noted that whereas density has almost no effect on red wines (its mode estimate is very close to 0), it has the strongest negative effect on white wines compared to other explanatory variables). Similarly, residual sugar has almost no effect on quality of red wine, but it has the largest positive effect for white wines. On the other hand, total sulfur dioxide has almost no effect on white wines but a strong negative effect on red wines. All other variables have mode estimates of corresponding β parameters close to zero, which means that they have insignificant impact on wine score. Lastly, it should be noted that for most of variables posterior distribution of β parameters has wider shape than for white wines (except for density, alcohol and residual sugar). This means that there is more uncertainty over β coefficients of red wines then of whites.

Now that we have estimated posterior distribution of model parameters, we will estimate posterior predictive distribution for some wines. The posterior predictive distribution represents the distribution of our response variable ('quality') integrated over the whole posterior distribution of model parameters, β and σ in particular, as we assumed that quality is distributed Normally with likelihood described above in this memo. We took three wines for checking the accuracy of prediction of chosen model, one wine with low quality score (top plot), one with average quality score (middle plot) and one with high score (bottom plot). Red vertical lines represent observed actual score of the wine.



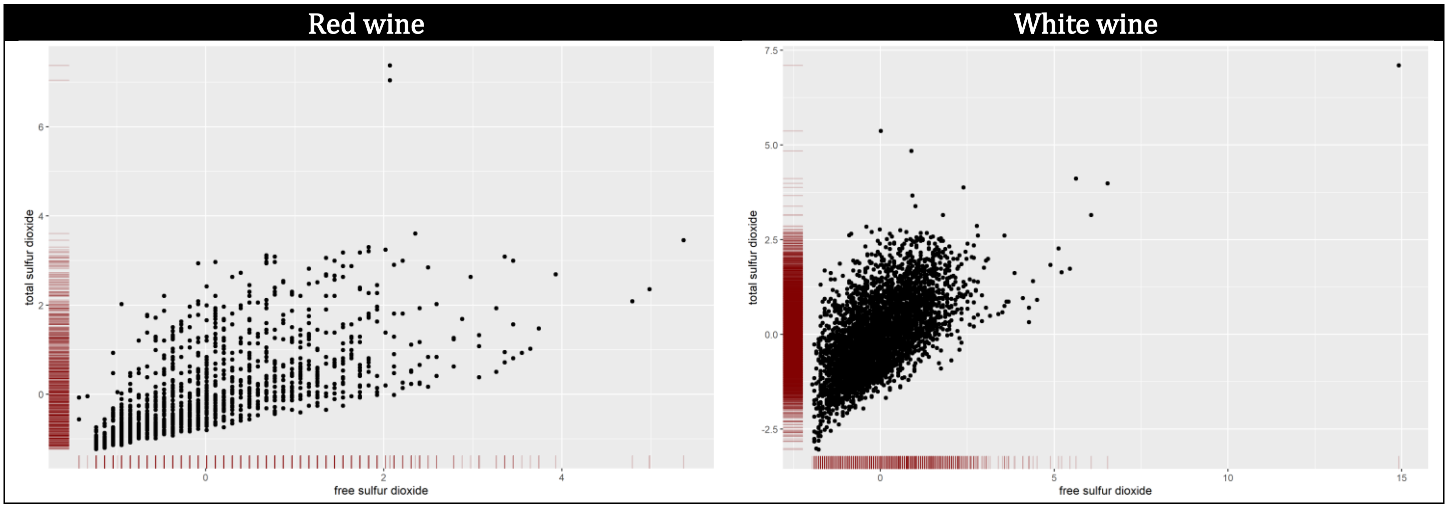


As it can be seen from the plots, pdf function peaks between scores 5 and 6 for both types of the wines. This makes the model good at prediction of average quality scores and rather poor predictor of quality scores at lower and higher ends.

Normal response model – transformed variables

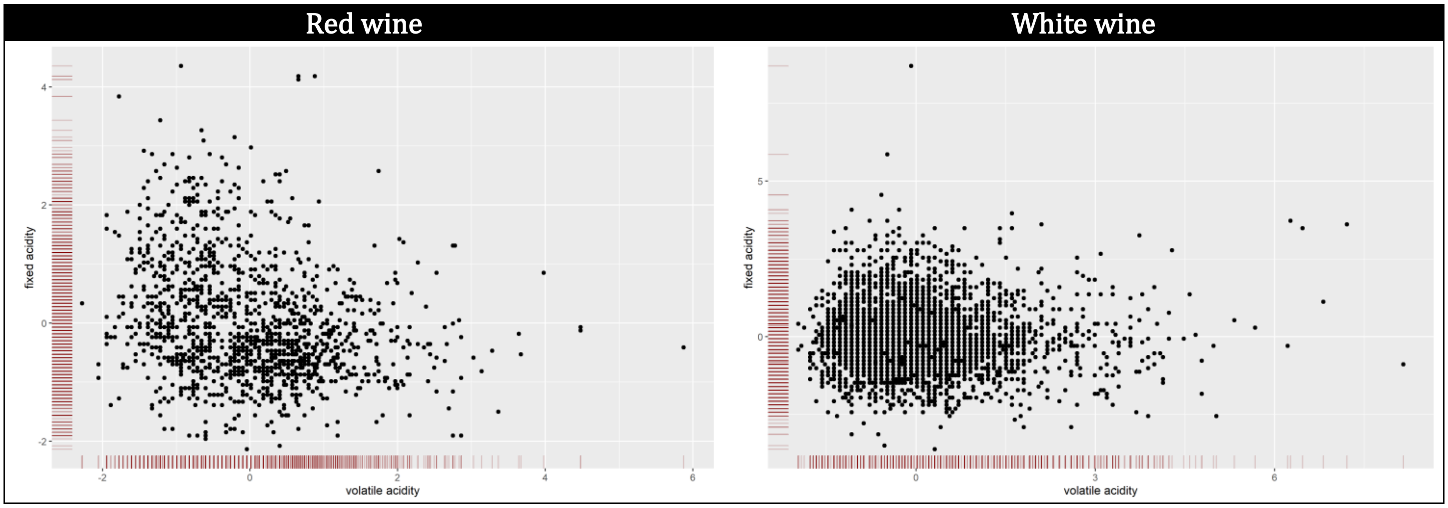
Next step is to consider transformation of some variables. In this project, we performed transformation of the following variables:

1. Proportion of free sulfur dioxide in total. We can see that free and total sulfur dioxide have high positive correlation in both wine datasets (0.7 for red and 0.6 for white wine). Scatter plots below show that the correlation is visible for all range of values but the variance between free and total sulfur grows with increase of sulfur amount.



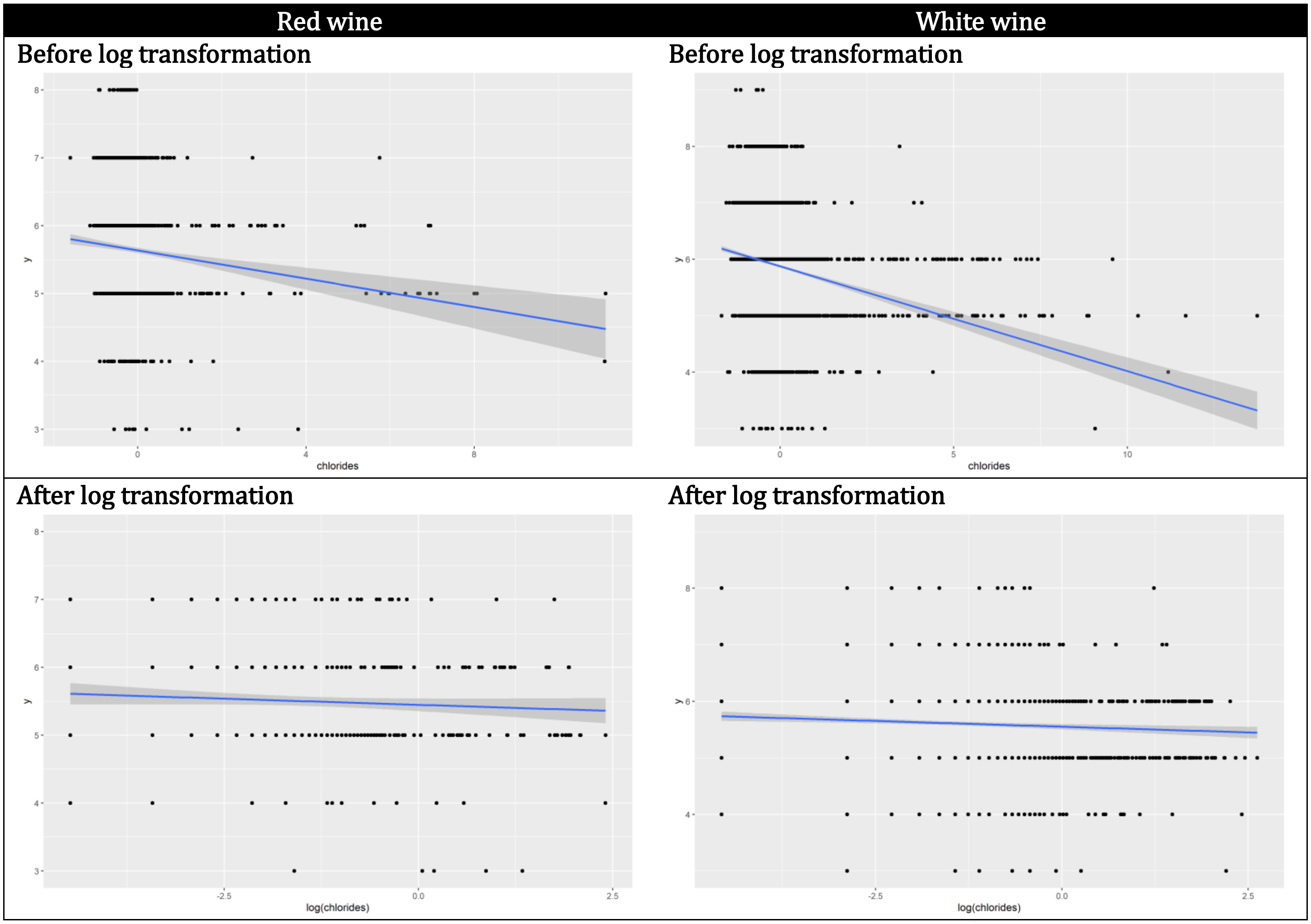
So, it may be worth to build model that includes the proportion of free sulfur dioxide in total.

2. Ratio of volatile to fixed acidity. As explained above volatile acidity have negative effect on wine quality score whereas fixed acidity has mall positive effect. Scatterplots show almost no correlation between these variables, but nevertheless we will have a look on the effect on the quality score from the ratio of these two variables.


Red wine

White wine

3. Log transformation of 'chlorides' variable. There are significant outliers in chlorides that impact correlation of this variable with quality of the wine (see below). We will performed log transformation of 'chlorides' in order to normalize that distribution of this variable.


Red wine

White wine

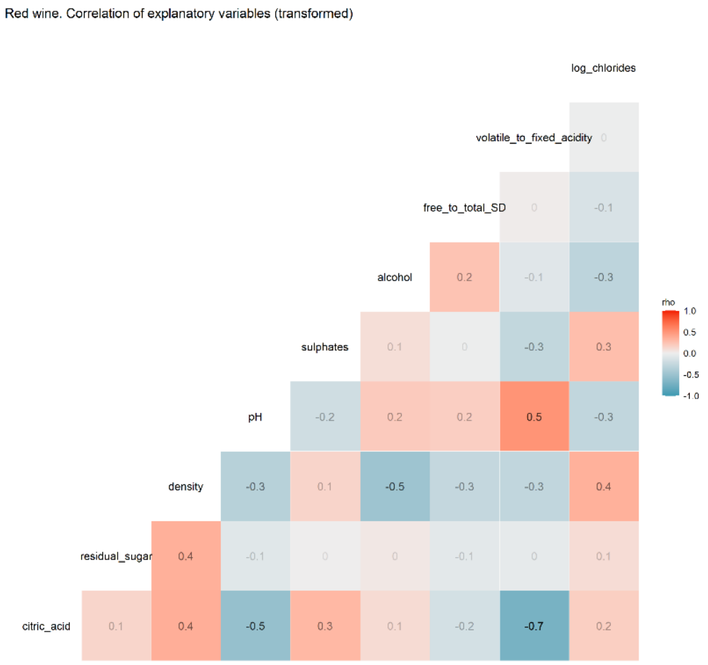
Before log transformation

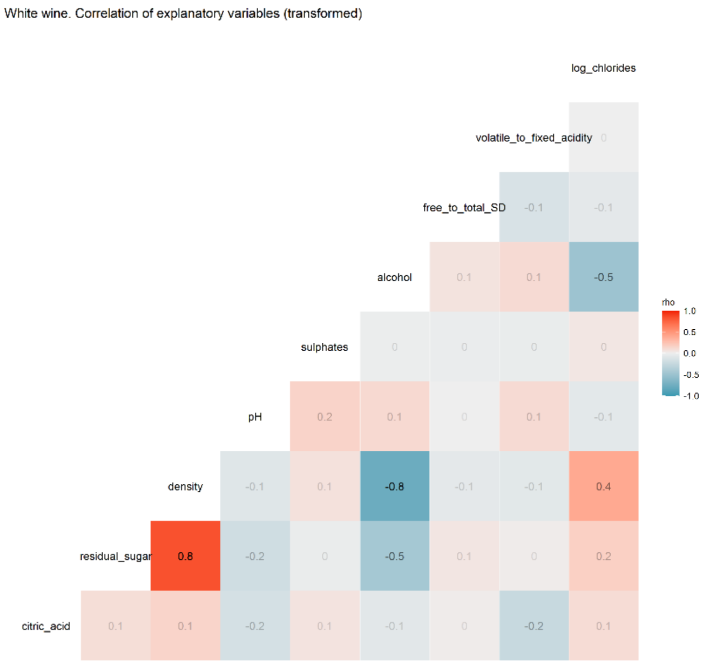
Before log transformation

After log transformation

After log transformation

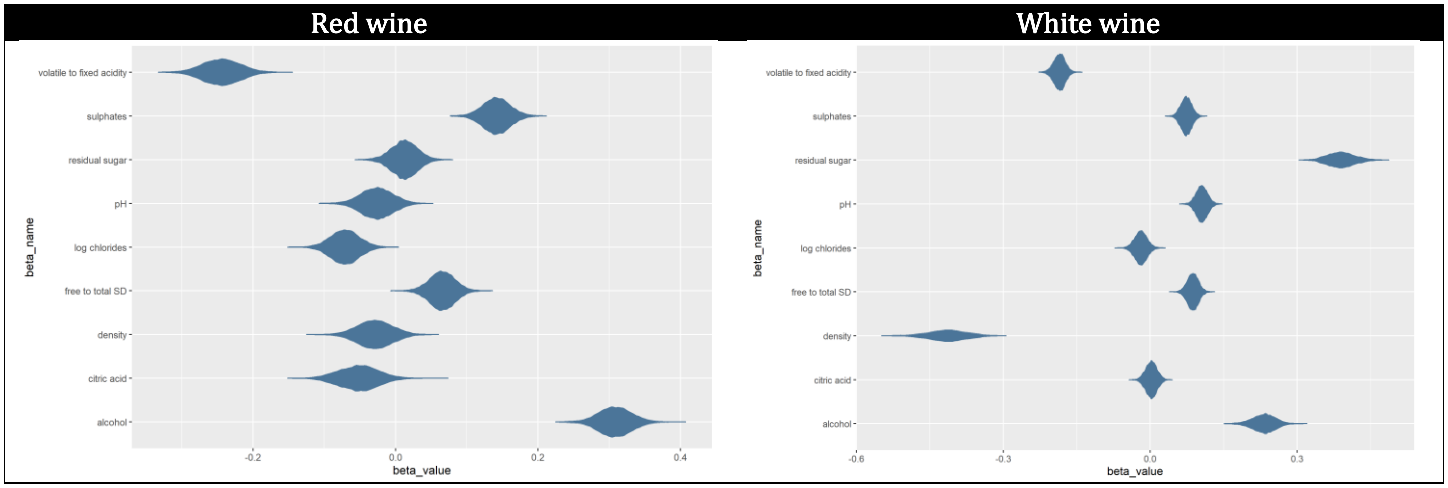
We expended set of explanatory variables with new variables explained above. At the same time, as these variables are derived from original explanatory variables, we excluded them from the analysis. Resulting datasets show the following correlation between transformed variables.





In white wines dataset the only significant correlation with absolute value above 0.5 is between density and alcohol and residual sugar. In red wine dataset ration of volatile to fixed acidity still has strong correlation with citric acid and pH, which is expected but at the same time there is almost no such correlation in white wine dataset.

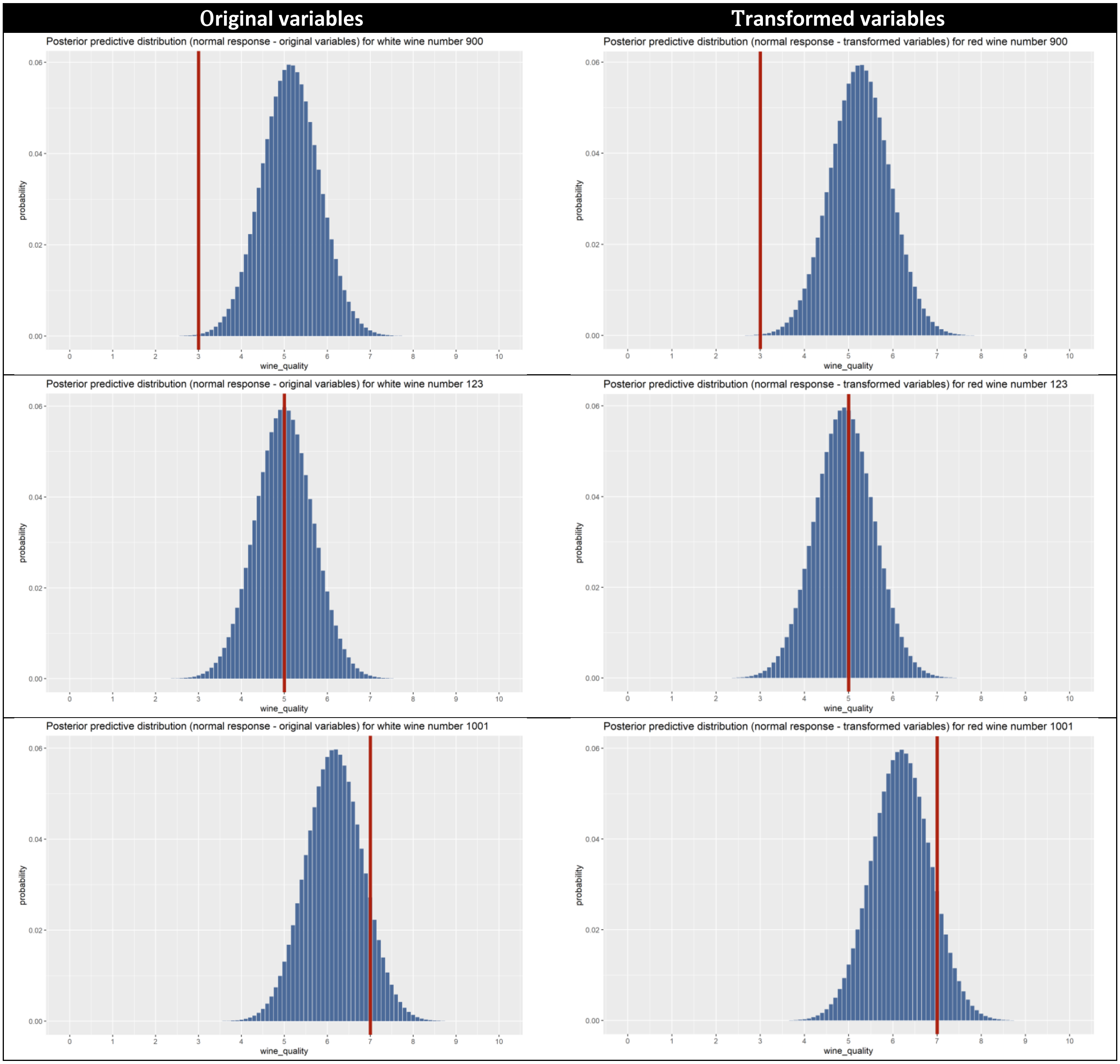
Now we will reperform sampling using Metropolis Hastings method same as discussed above. Posterior distribution of β coefficients for transformed explanatory variables is the following.



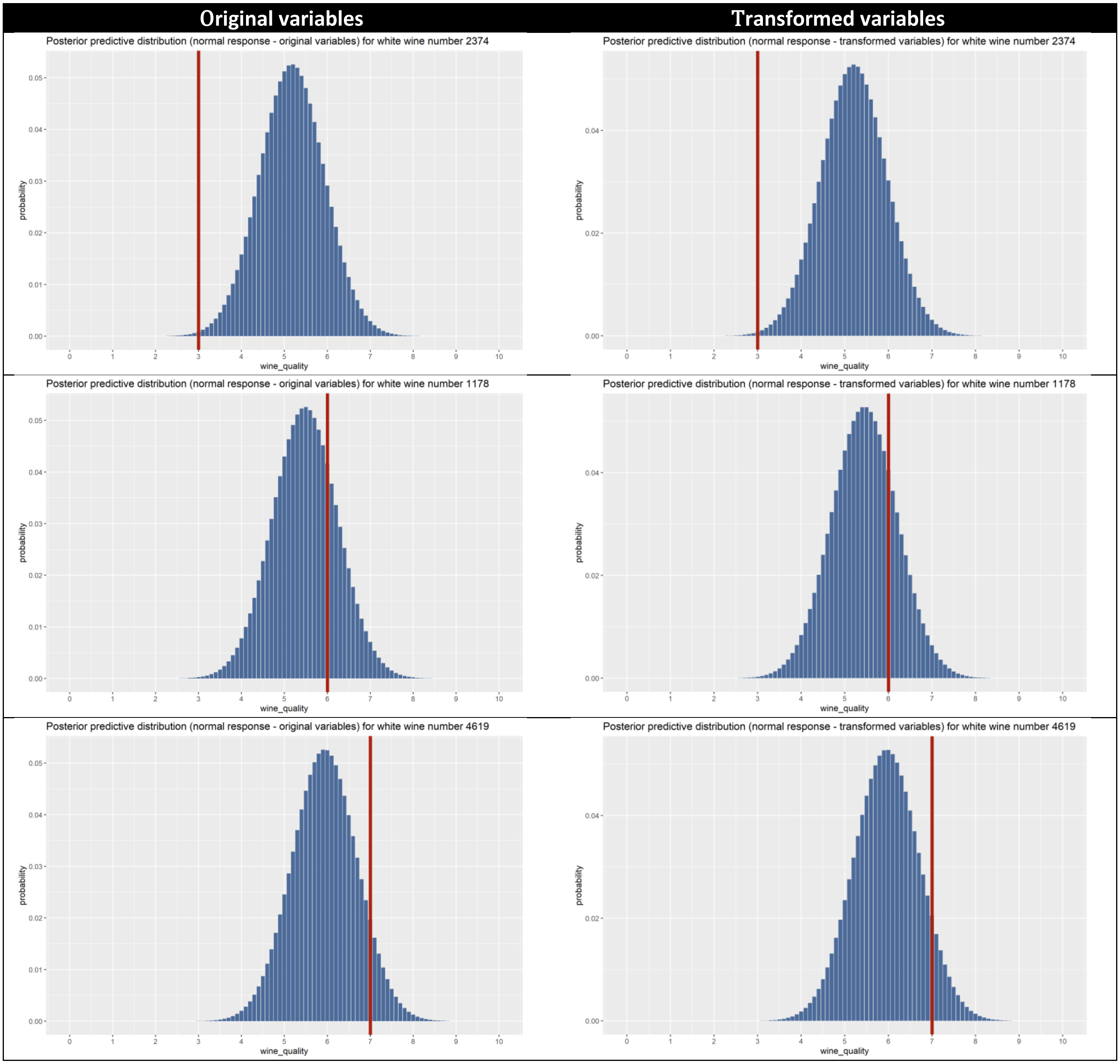
For both wines ratio of volatile to fixed acidity β still has large negative mode estimate, which means negative effect on the wine quality score. Log chlorides has almost no effect and free to total sulfur dioxide ration has small positive effect. In general, posterior distribution of β of transformed variables resembles that of original variables in many senses.

Accordingly, posterior predictive distribution of is not significantly different from original dataset. Below plots compare posterior predictive distribution of quality score with model with original variables with posterior predictive distribution with transformed variables.

A. Red wines



B. White wines



Plots above show that posterior predictive distribution is almost identical. It means that transformation of variables did not add value to the model. Maybe more complex transformations that are not covered in this memo will give better results.

Multinomial response model

Normal models above assume that 'quality' is a continuous response variable, which is not true as the response can take only discrete values. Therefore, in order to incorporate this property of the response variable we will introduce another model where the 'quality' is distributed in accordance with Multinomial distribution.

In multinomial experiment an experiment consists of n repeated trials and each trial has discrete number k of possible outcomes and each outcome has a particular probability θ_i to occur so that sum of all probabilities θ for all possible outcomes is equal to 1. If to apply, the model to red wine datasets then each row in the dataset will represent 1 trial where 'quality score' can take 1 out of 6 possible scores (from 3 to 8 including), where each score has a probability of occurrence θ. We can rewrite the model as follows. Let y_i be the categorical dependent variable ('quality') for observation i which takes an integer values j (from 3 to 8).

$$y_i \sim Multinomial(y_i|\theta_{\mathrm{ij}})$$

So, the question here is what will be the value of θ. θ can be estimated using SoftMax function, also called normalized exponential function. The SoftMax function takes as input a vector z of K real numbers and normalizes it into a probability distribution consisting of K probabilities proportional to the exponentials of the input numbers. That is, prior to applying SoftMax, some vector components could be negative, or greater than one; and might not sum to 1; but after applying SoftMax, each component will be in the interval (0, 1), and the components will add up to 1, so that they can be interpreted as probabilities. Furthermore, the larger input components will correspond to larger probabilities.

With regard to the wine datasets the SoftMax function will be as follows:

$$\theta_{ij} = \frac{e^{X_i \beta_j}}{\sum_{k=1}^{K} e^{X_i \beta_j}}$$

Where X_i is the explanatory variable for wine y_i and β_j is beta parameters for wine of a particular wine score out of K possible (i.e. 6 in case of red wine data set). In this project, prior β parameter is assumed to be Normally distributed as follows:

$$\beta \mid \sigma^2 \ \sim Normal(0, \sigma^2)$$

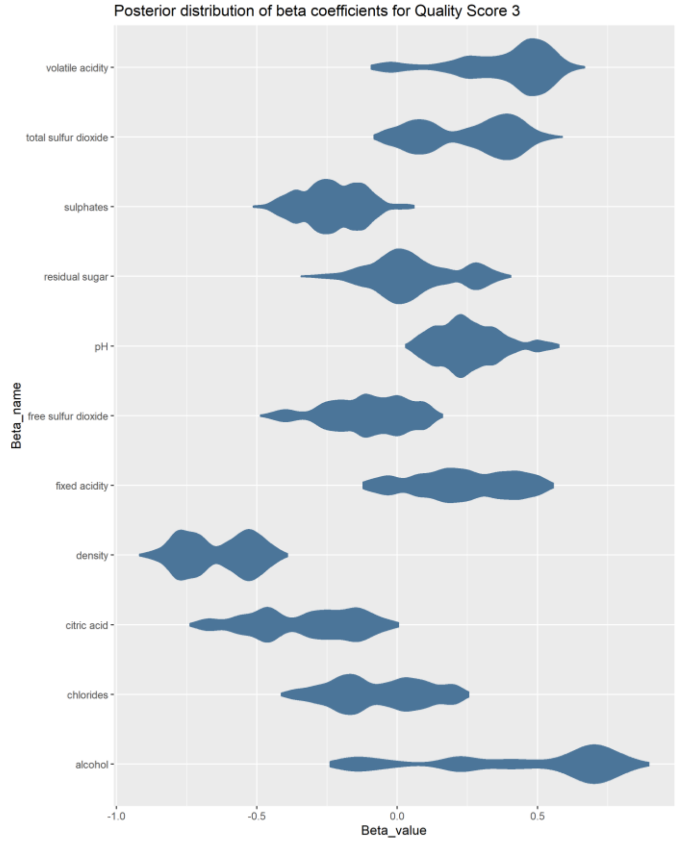
$$\sigma^2 \ \sim Inv - Gamma(a, b)$$

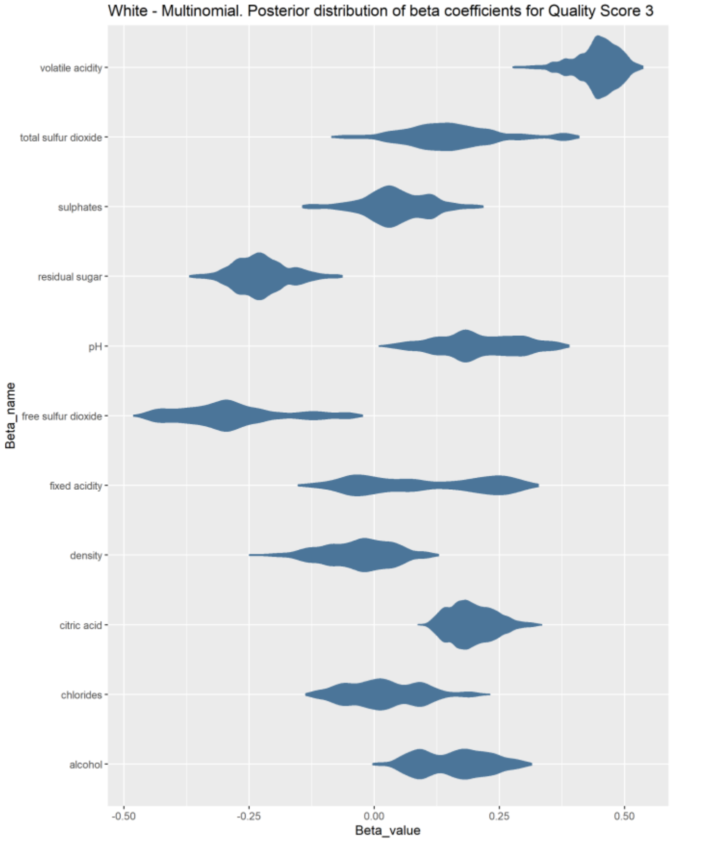
σ^2 has the same prior distribution and same prior hyperparameters as in Normal response model.

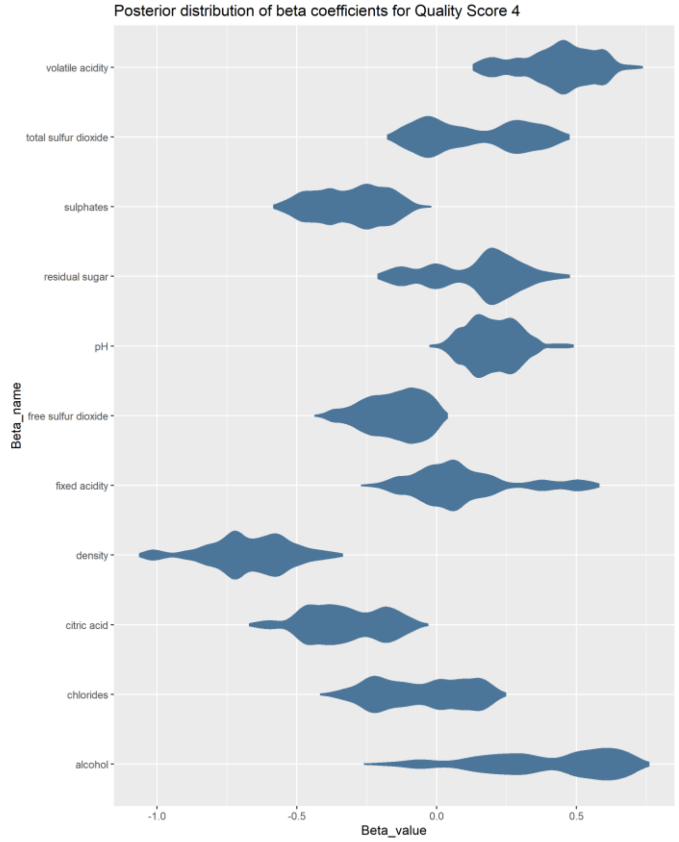
It should be noted that in Multinomial model each outcome will have its own β parameters and also in order to technically implement the model in R code vector of response variable 'quality' of length n was transformed matrix with n rows and K columns, where n is the number of observations and K is the number of distinct outcomes. Each row of this matrix contains only 1 value of **one** and K-1 **zeroes.** Position of **one** indicates the outcome. For example, as mentioned above the red dataset contains 6 distinct quality scores (3, 4, 5, 6, 7, 8) so if a row i of the matrix looks as follows (0, 0, 0, 1, 0) indicates the i observation of has quality score of 7. This is done in order to properly implement multinomial model where the vector of number of distinct outcomes needs to be the same length as the vector of probabilities θ of each outcome. Each wine quality represent a single separate trial where only one particular quality score is observed and no other scores.

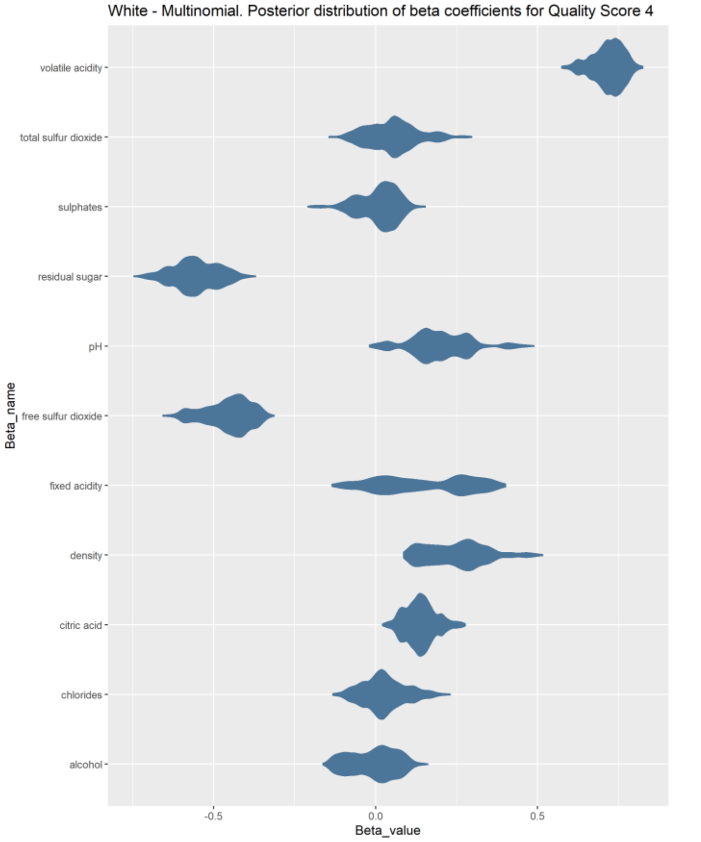
After defining the model, let's derive posterior distribution of model parameters using Metropolis-Hastings sampling method. In case of Multinomial model, we took in total 10,000 samples for each wine type and then removed first 5,000 samples leaving 5,000 samples of representing posterior distribution of model parameters β and σ^2. The violin plot below shows posterior distribution of β for red and white wine. Same as for Normal model, I made 3 iterations 10,000 samples each iterations in order to check convergence so that posterior distribution is the same for each draw. The draws are saved in the folder draws that is submitted together with the other files of the project. The comparison of posterior distribution is presented in the **Appendix 2**. Posterior distribution of parameters for each quality score of Draw # is presented below:

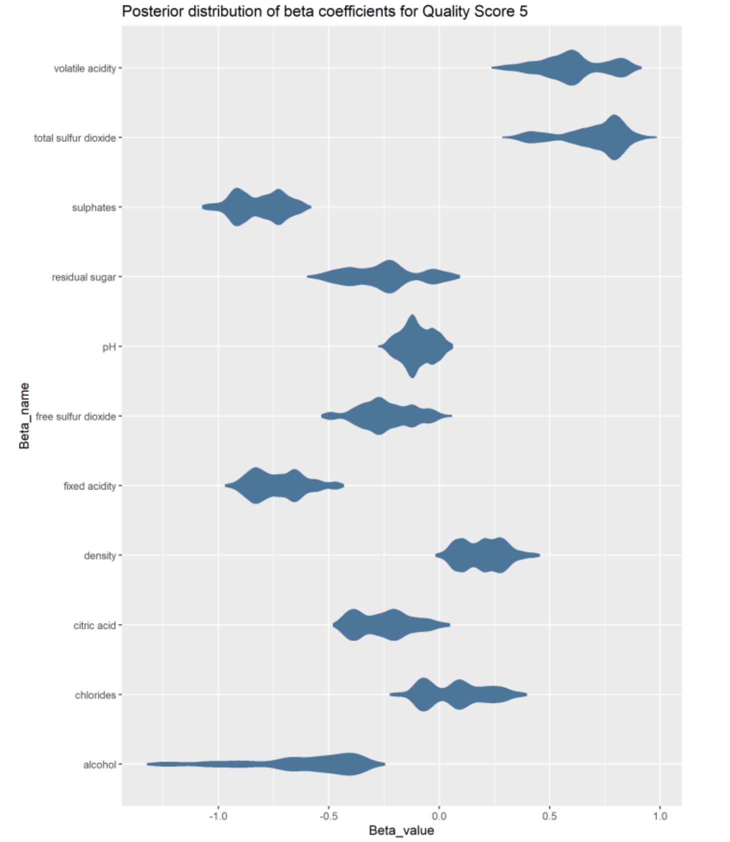
Red wine	White wine

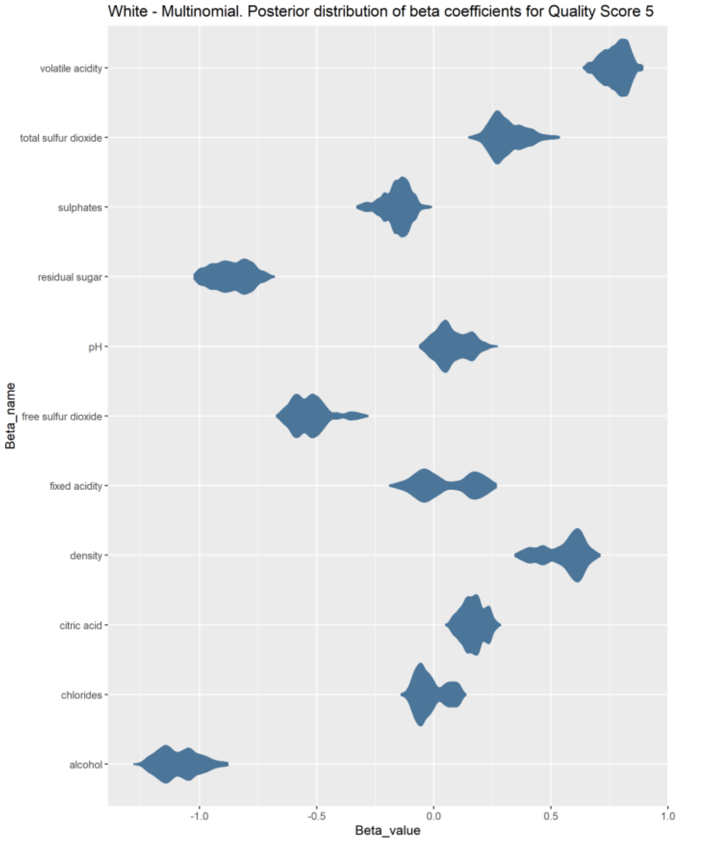

Posterior distribution of beta coefficients for Quality Score 3

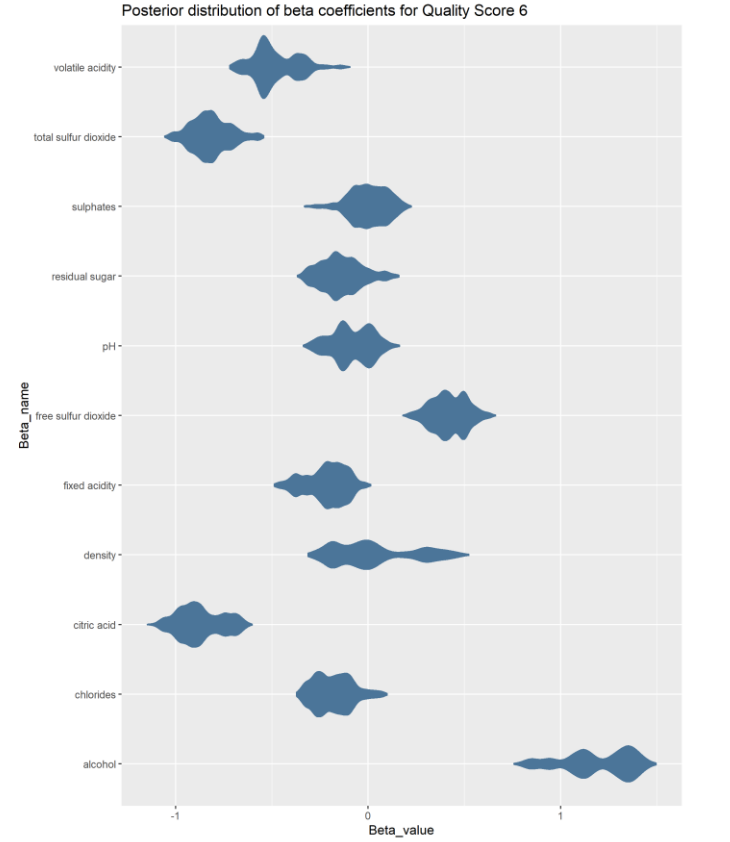

White - Multinomial. Posterior distribution of beta coefficients for Quality Score 3

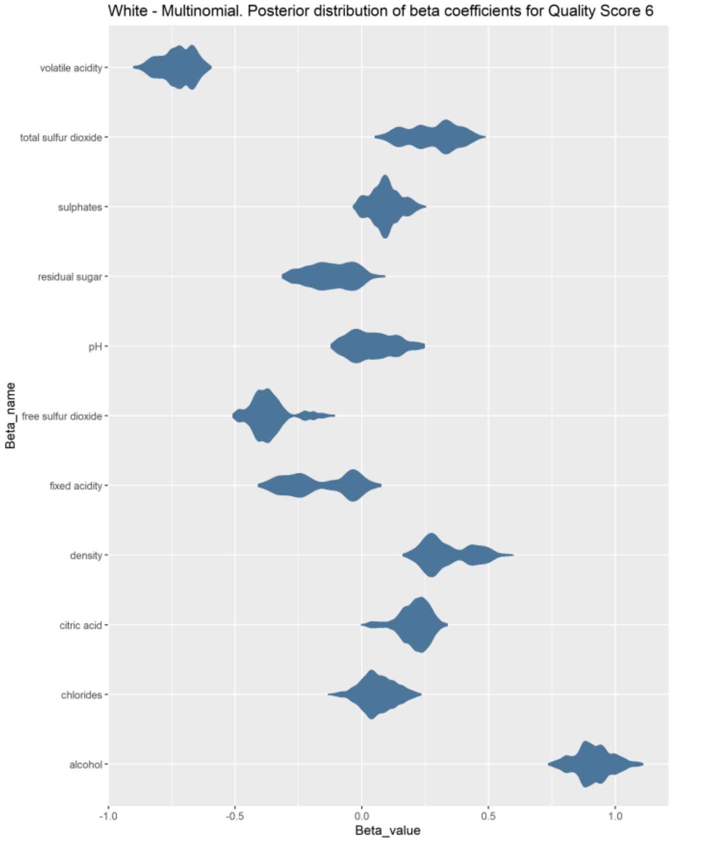

Posterior distribution of beta coefficients for Quality Score 4

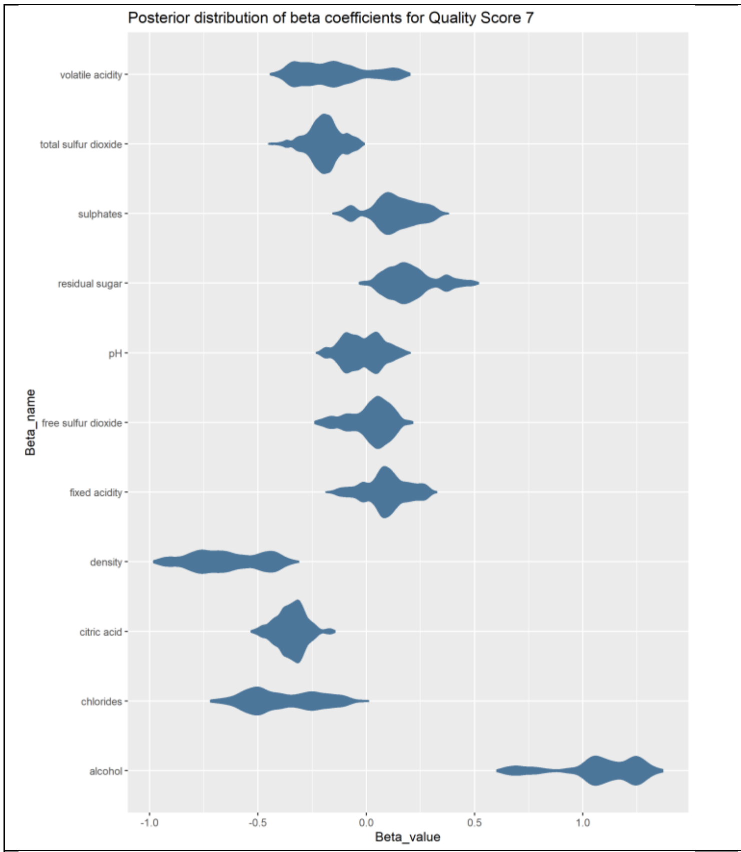

White - Multinomial. Posterior distribution of beta coefficients for Quality Score 4


Posterior distribution of beta coefficients for Quality Score 5
Beta_name
volatile acidity
total sulfur dioxide
sulphates
residual sugar
pH
free sulfur dioxide
fixed acidity
density
citric acid
chlorides
alcohol
-1.0
-0.5
0.0
0.5
1.0
Beta_value

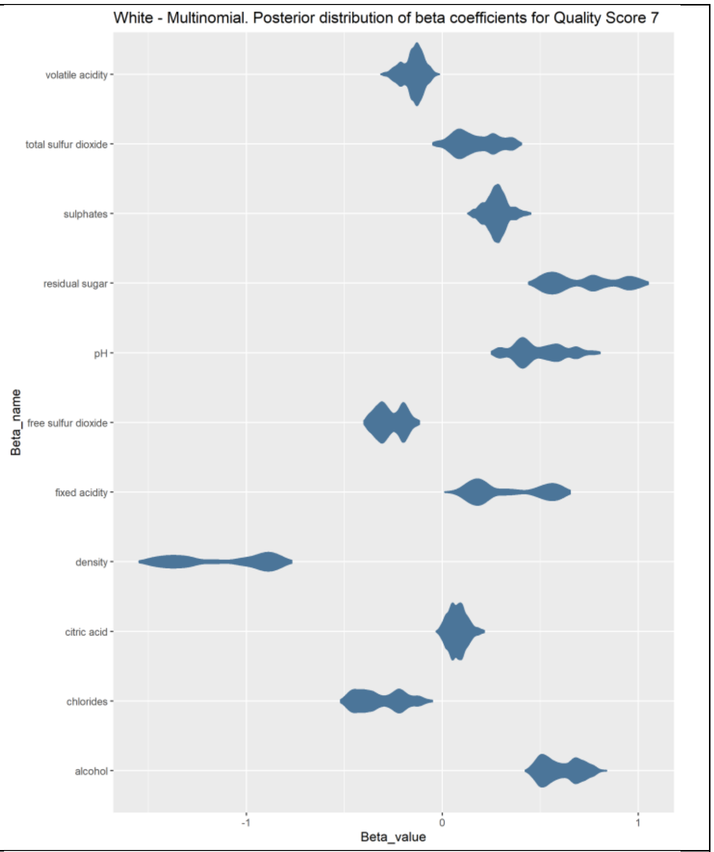

White - Multinomial. Posterior distribution of beta coefficients for Quality Score 5
Beta_name
volatile acidity
total sulfur dioxide
sulphates
residual sugar
pH
free sulfur dioxide
fixed acidity
density
citric acid
chlorides
alcohol
-1.0
-0.5
0.0
0.5
1.0
Beta_value


Posterior distribution of beta coefficients for Quality Score 6
Beta_name
volatile acidity
total sulfur dioxide
sulphates
residual sugar
pH
free sulfur dioxide
fixed acidity
density
citric acid
chlorides
alcohol
-1
0
1
Beta_value

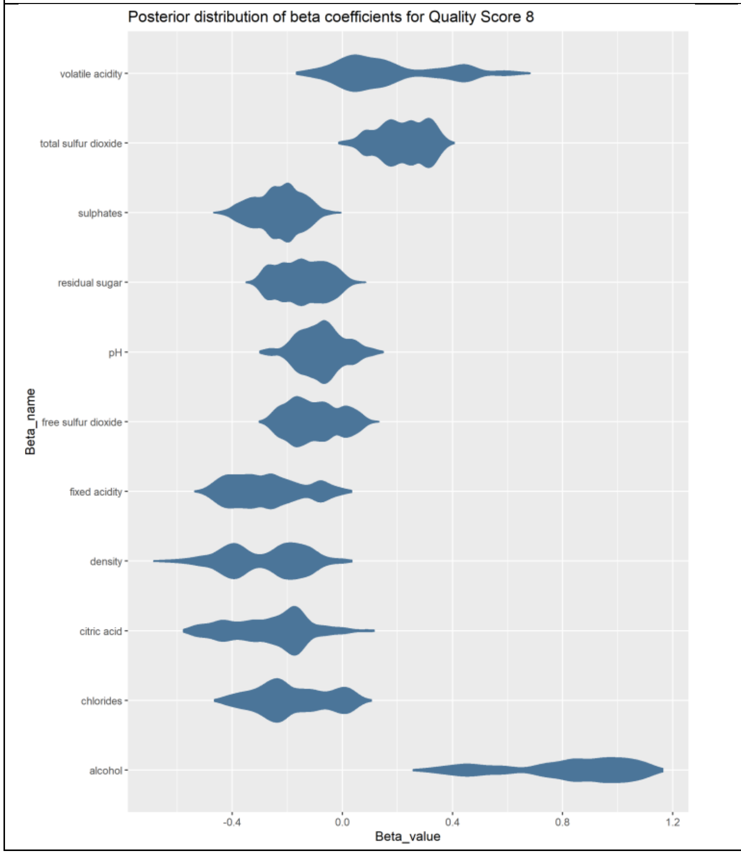

White - Multinomial. Posterior distribution of beta coefficients for Quality Score 6
Beta_name
volatile acidity
total sulfur dioxide
sulphates
residual sugar
pH
free sulfur dioxide
fixed acidity
density
citric acid
chlorides
alcohol
-1.0
-0.5
0.0
0.5
1.0
Beta_value



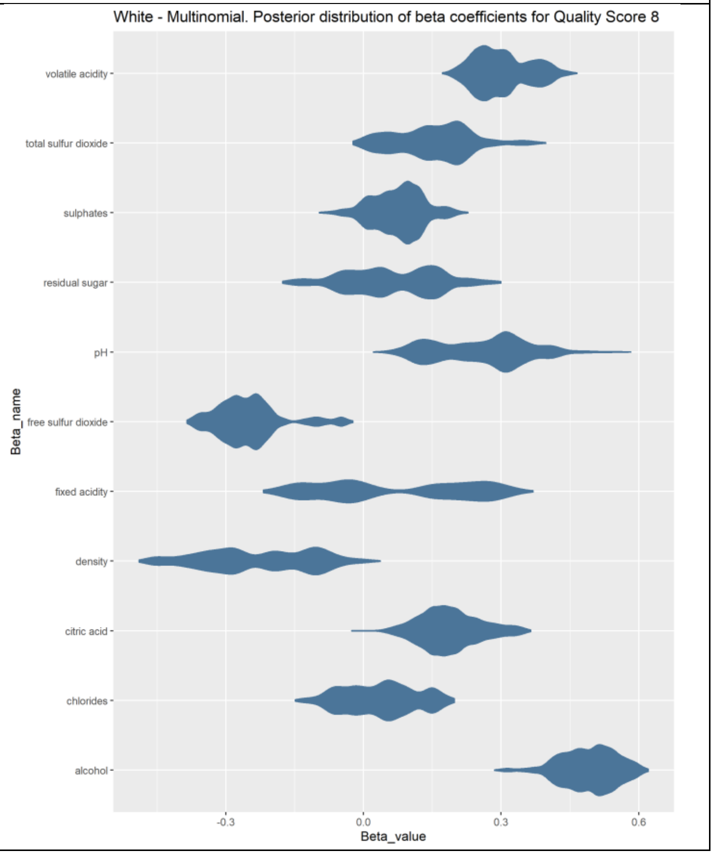
Posterior distribution of beta coefficients for Quality Score 7



White - Multinomial. Posterior distribution of beta coefficients for Quality Score 7

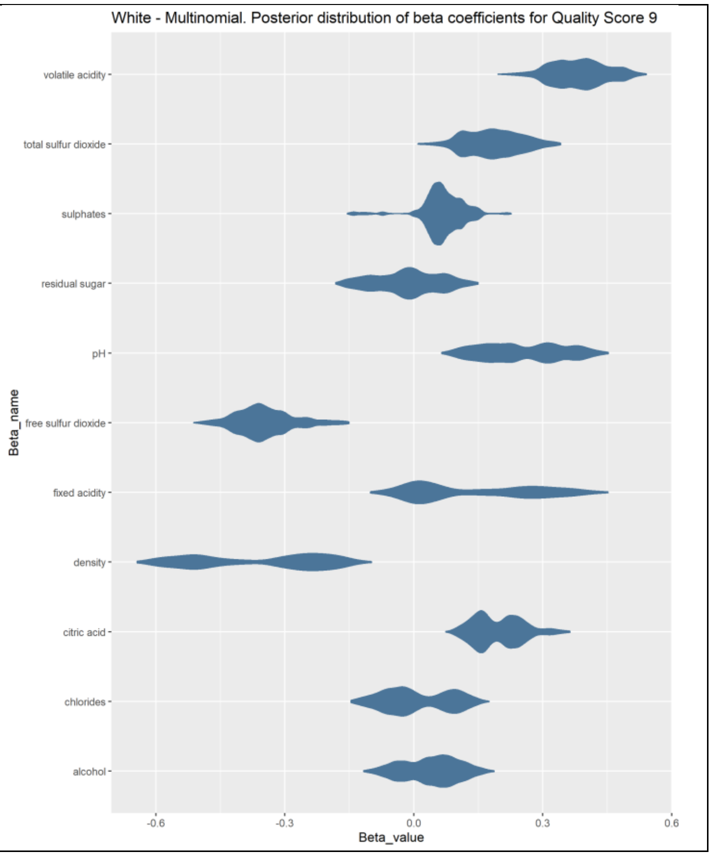


Posterior distribution of beta coefficients for Quality Score 8



White - Multinomial. Posterior distribution of beta coefficients for Quality Score 8

There are no wines with 9 score in red wines data set



The plots above show that distribution of beta parameters is wide indicating lack of convergence. Also, comparison of violin plots from different draws in the **Appendix 2** shows that mode estimate of beta coefficients can be different for each sampling iteration that also may indicate that samples did not converge. At the same time, posterior predictive distributions for each of 3 sampling iterations of red and white wine datasets presented in the **Appendix 3** are almost identical to each other regardless difference in posterior distribution of beta coefficients. Difference between posterior distribution of beta coefficient together with almost identical posterior predictive distribution results from more complex link between response variable and exploratory variables and beta coefficients via SoftMax function. **Similarity of posterior predictive distributions using different samples is a good indication of robustness of the model, so we will not discard this model and analyze it further.**

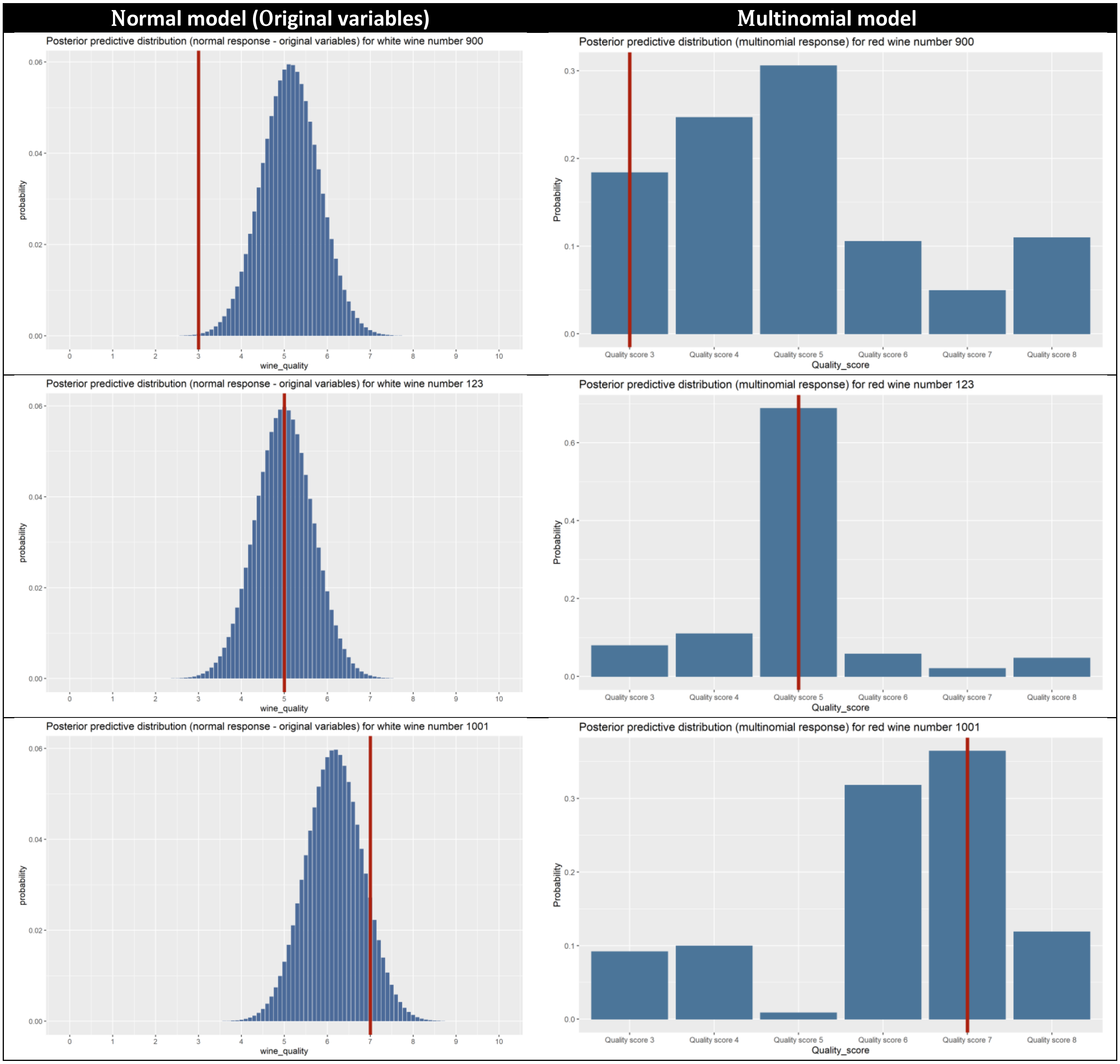
Although posterior distribution of betas is different between different draws, some similarities are noticeable:

- Unlike Normal model, mode estimate of posterior beta for alcohol is negative both for red and white wines of quality 5 in Multinomial model. For other quality scores the mode estimate is positive, especially for wines with score 6 and 7. This can be interpreted as follows: increasing alcohol content in the wine (with other exploratory variables fixed) decreases the probability of this wine to have the quality score 5.
- Posterior beta of volatile acidity has positive mode estimate for wines with lower scores and negative with wines with higher scores. As explained above, volatile acidity is associated with unpleasant vinegary taste so it is expected that wines with higher volatile acidity will get lower quality scores, hence the positive mode estimate. Unexpected positive mode estimate of volatile acidity beta for white wines with score 9 in the end of the above plot is due to lack of convergence and the result is not consistent among other draws as presented in the **Appendix 2**.

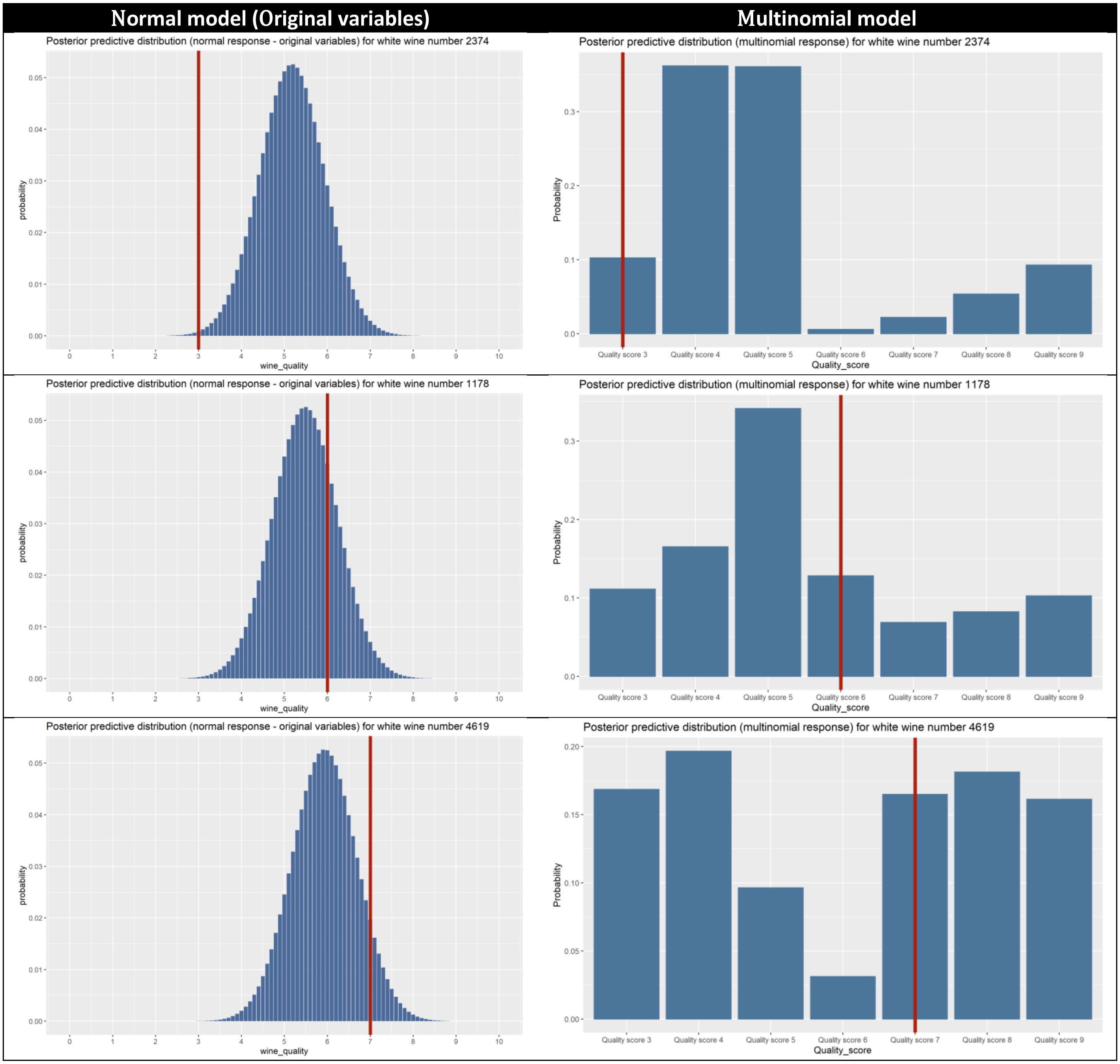
There is not much more can be taken from posterior distribution of betas alone, so will calculate posterior predictive distribution and compare its accuracy of prediction of quality with that of Normal model. In order to calculate posterior predictive distribution of a particular wine, we need to calculate probabilities θ of this wine to get each of the possible score. Probabilities θ can be calculated by integrating SoftMax function with observable explanatory variables of this wine over the whole range posterior distribution of beta coefficients of each possible quality score.

Resulting vector of θ of length of a number of all possible scores needs to be then plugged in Multinomial probability distribution functions. But as explained above, each observation (row) in wine datasets represent only 1 trial with wine taking only score and not taking other scores. This already reflected in vector of θ. Comparison of the posterior predictive distribution on Normal and Multinomial models is presented below:

A. Red wines



B. White wines

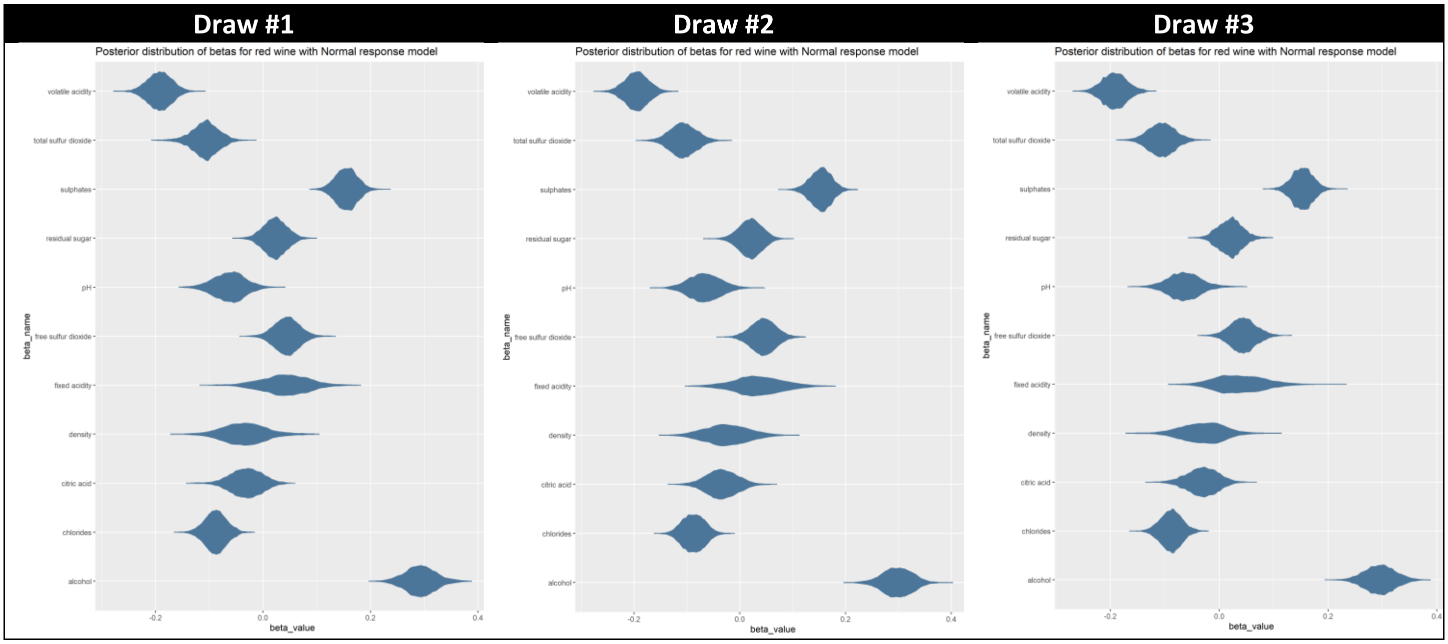


Conclusion

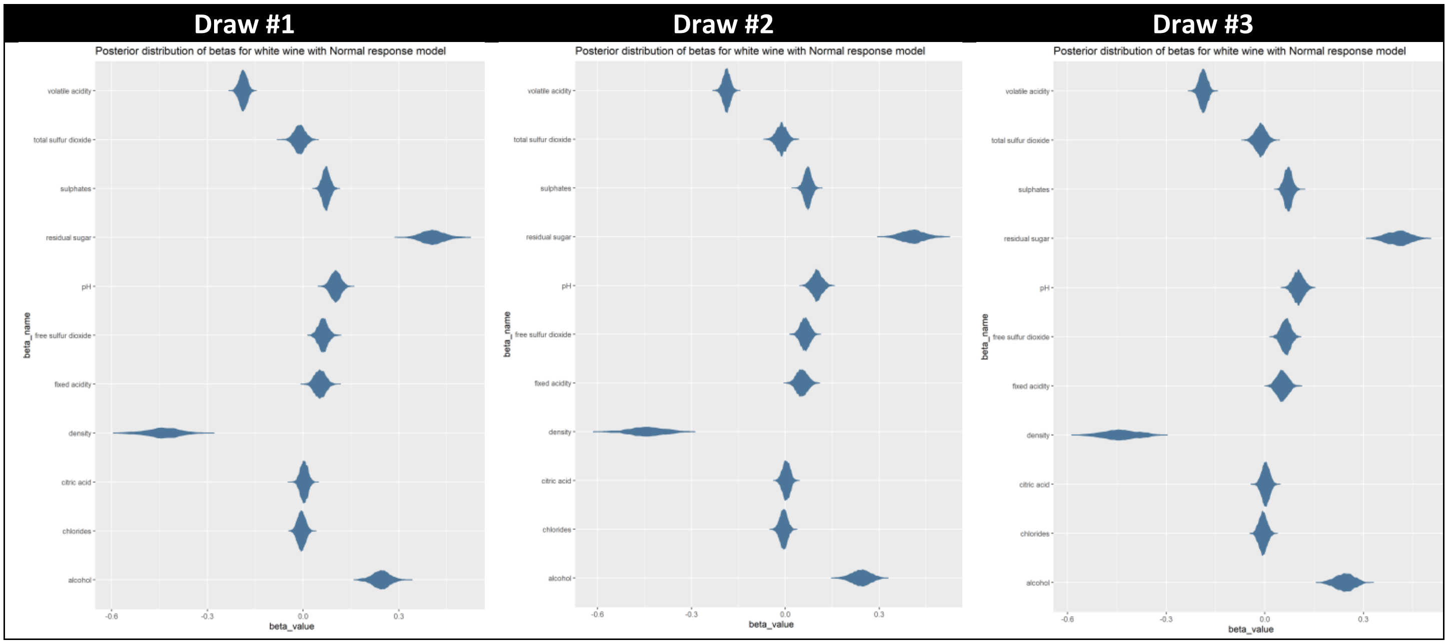
The drawback of the Normal model was that its posterior predictive distribution peaked around 5 and 6 so the Multinomial model poorly predicts lower and higher quality scores. As in case of Normal model, Multinomial response model predicted accurately wine of score 5 in red dataset. But at the same time, Multinomial model, although not very accurate, does better job at predicting quality scores at higher and lower end. For example, it predicted more accurately red wine with score 7.

APPENDIX 1. Posterior distribution of model parameters of model with Normal response

Posterior distribution of beta parameters for red wine after taking 3 samples 120,000 draws each. The distribution is almost identical.

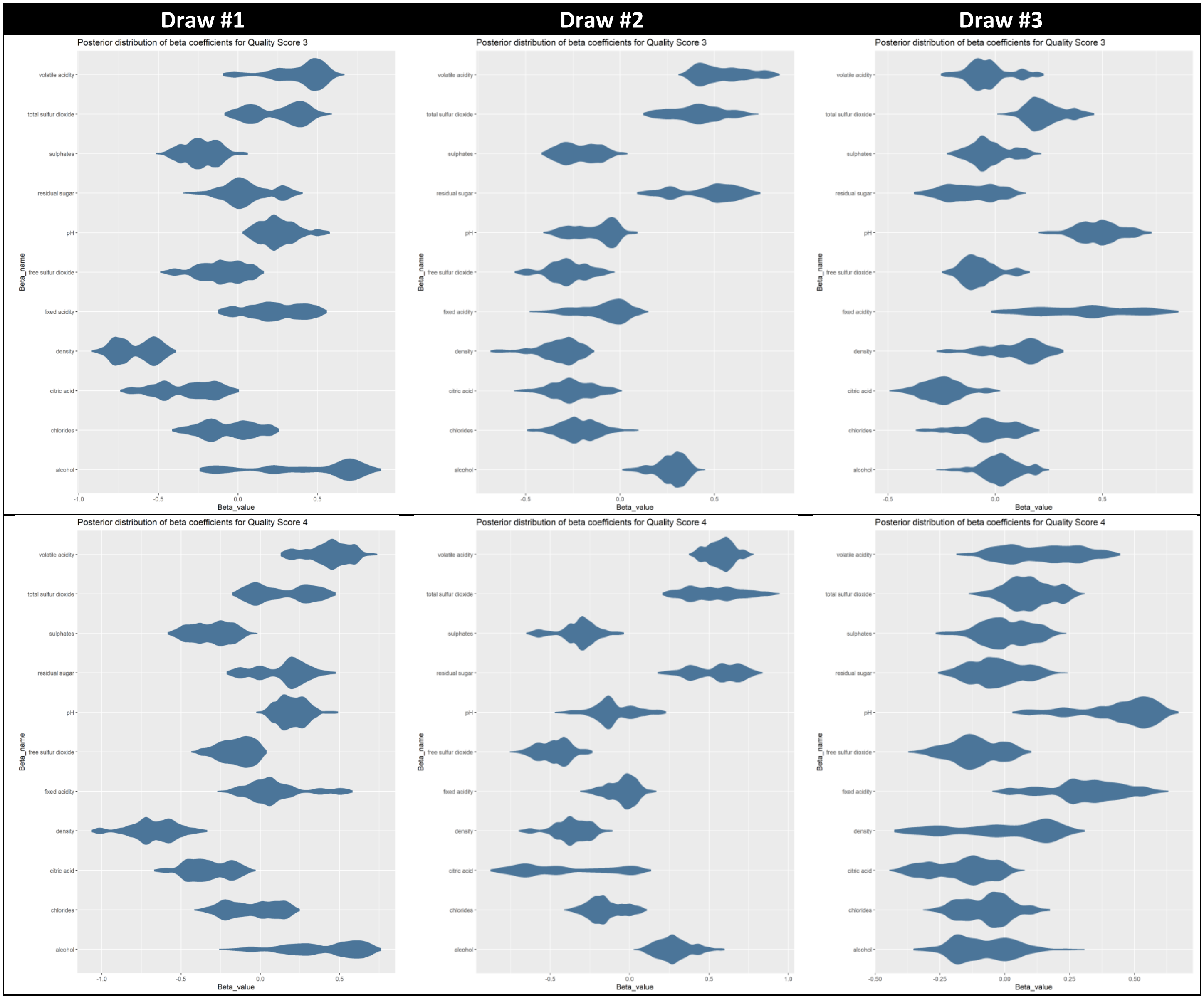


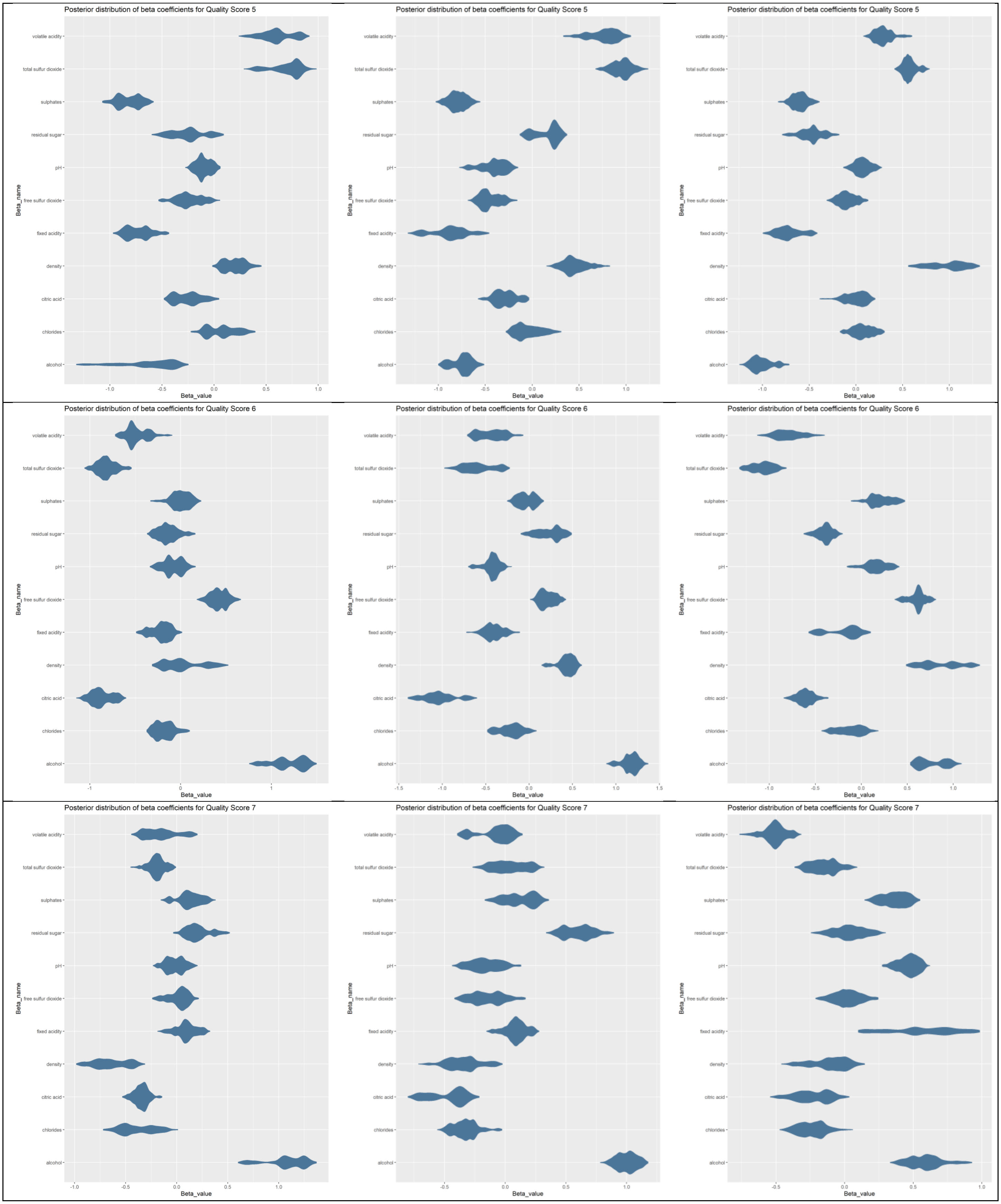
Posterior distribution of beta parameters for white wine after taking 3 samples 120,000 draws each. The distribution is almost identical.



APPENDIX 2. Posterior distribution of model parameters of model with Multinomial response

Posterior distribution of beta parameters for red wine after taking 3 samples 10,000 draws each. The distribution is wide for most of the betas and mode estimate is not stable.





Posterior distribution of beta coefficients for Quality Score 5

Posterior distribution of beta coefficients for Quality Score 5

Posterior distribution of beta coefficients for Quality Score 5

Posterior distribution of beta coefficients for Quality Score 6

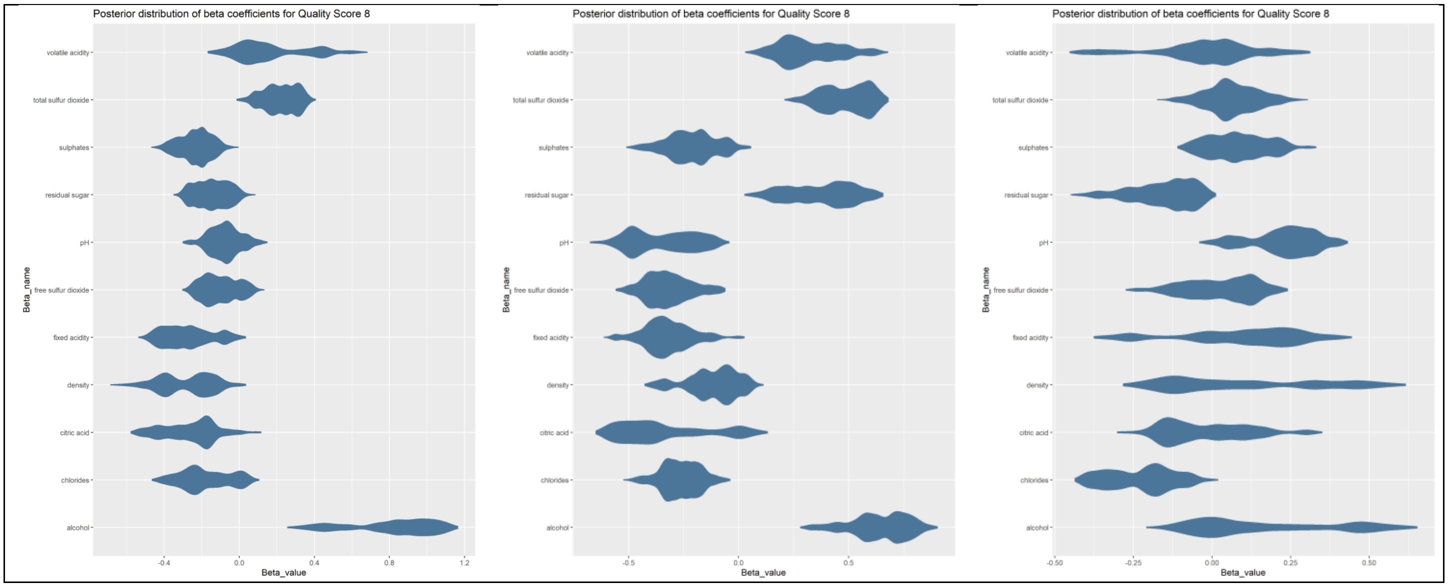
Posterior distribution of beta coefficients for Quality Score 6

Posterior distribution of beta coefficients for Quality Score 6

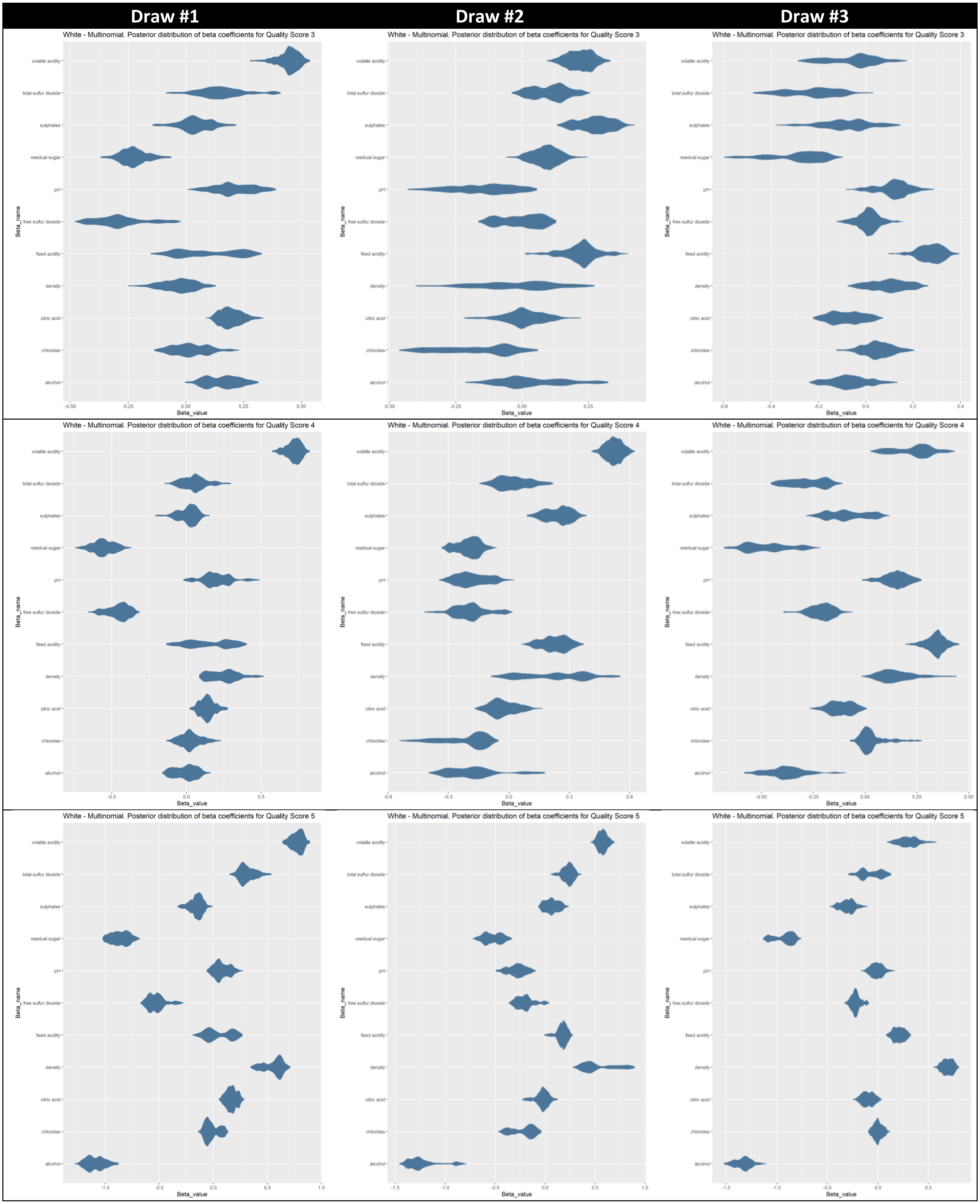
Posterior distribution of beta coefficients for Quality Score 7

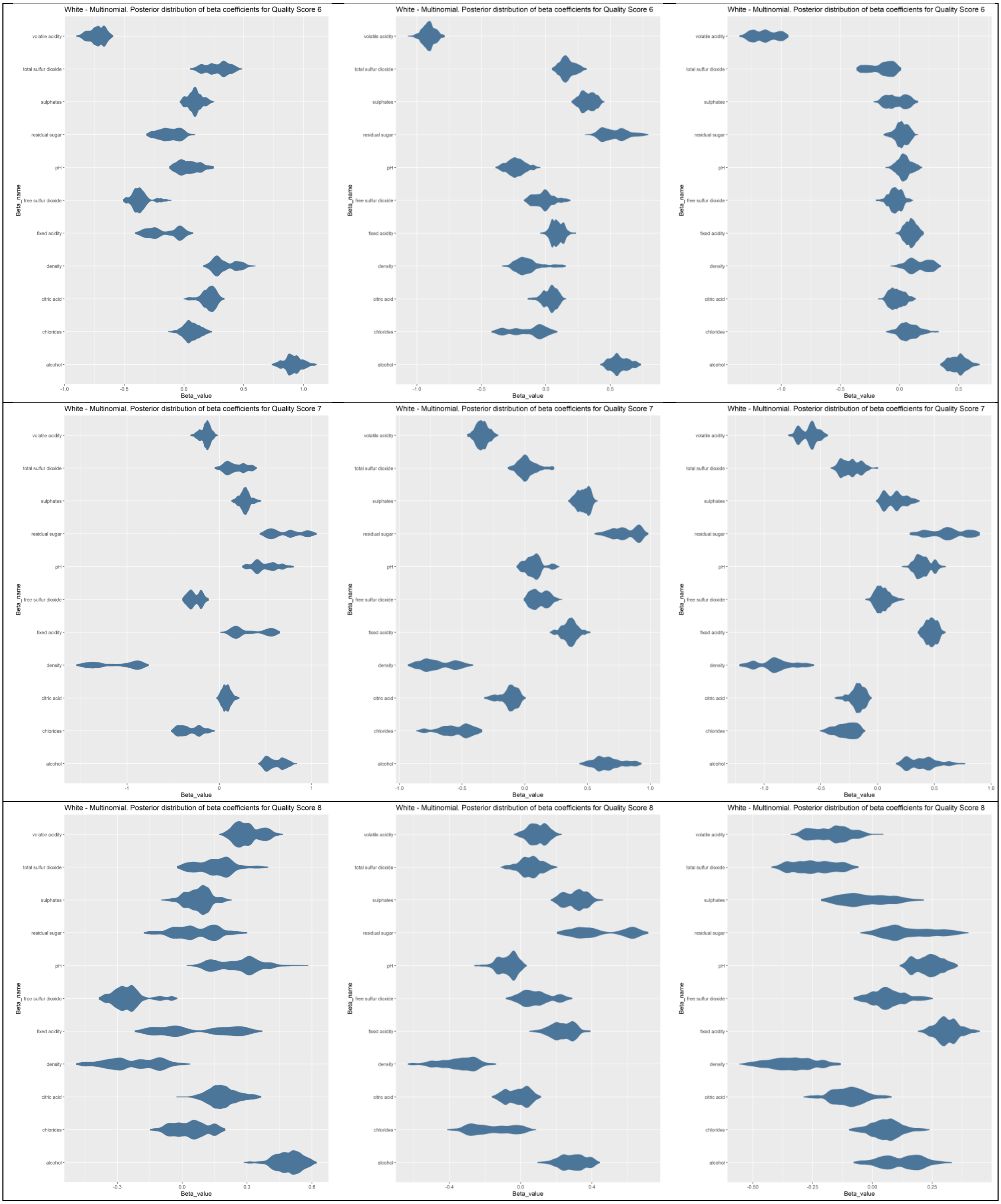
Posterior distribution of beta coefficients for Quality Score 7

Posterior distribution of beta coefficients for Quality Score 7


Posterior distribution of beta coefficients for Quality Score 8
Posterior distribution of beta coefficients for Quality Score 8
Posterior distribution of beta coefficients for Quality Score 8
Beta_name
volatile acidity
total sulfur dioxide
sulphates
residual sugar
pH
free sulfur dioxide
fixed acidity
density
citric acid
chlorides
alcohol
-0.4
0.0
0.4
0.8
1.2
-0.5
0.0
0.5
-0.50
-0.25
0.00
0.25
0.50
Beta_value

Posterior distribution of beta parameters for red wine after taking 3 samples 10,000 draws each. The distribution is wide for most of the betas and mode estimate is not stable.




White - Multinomial. Posterior distribution of beta coefficients for Quality Score 6

White - Multinomial. Posterior distribution of beta coefficients for Quality Score 6

White - Multinomial. Posterior distribution of beta coefficients for Quality Score 6

White - Multinomial. Posterior distribution of beta coefficients for Quality Score 7

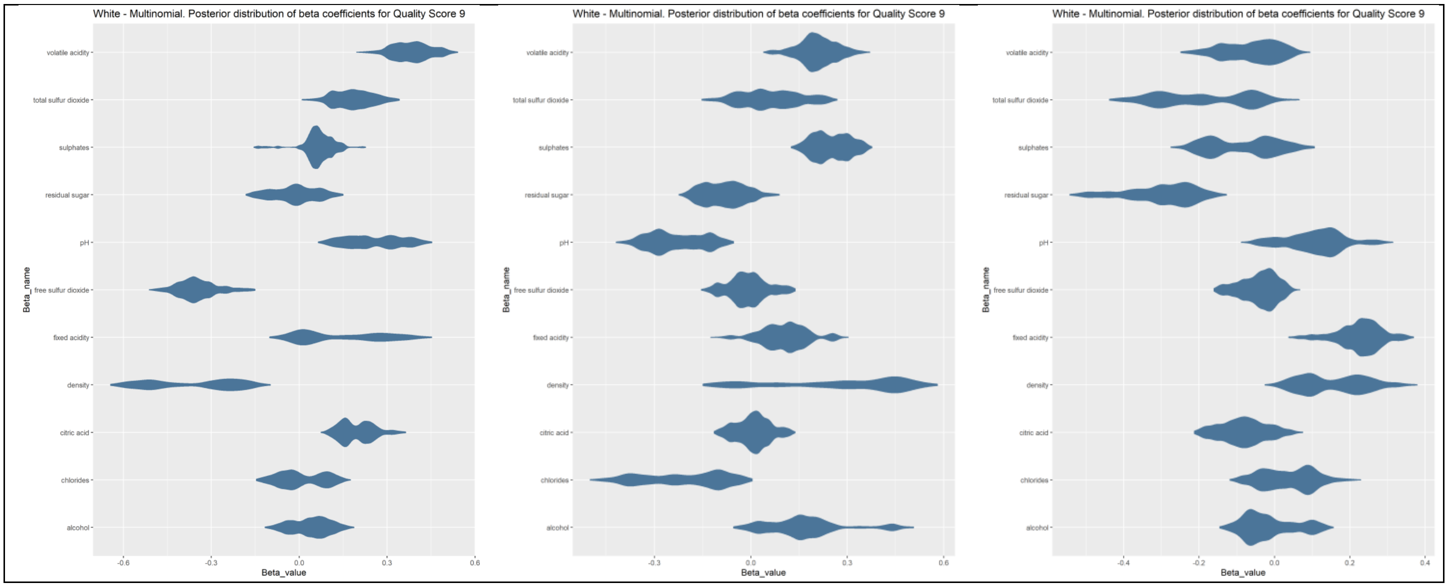
White - Multinomial. Posterior distribution of beta coefficients for Quality Score 7

White - Multinomial. Posterior distribution of beta coefficients for Quality Score 7

White - Multinomial. Posterior distribution of beta coefficients for Quality Score 8

White - Multinomial. Posterior distribution of beta coefficients for Quality Score 8

White - Multinomial. Posterior distribution of beta coefficients for Quality Score 8



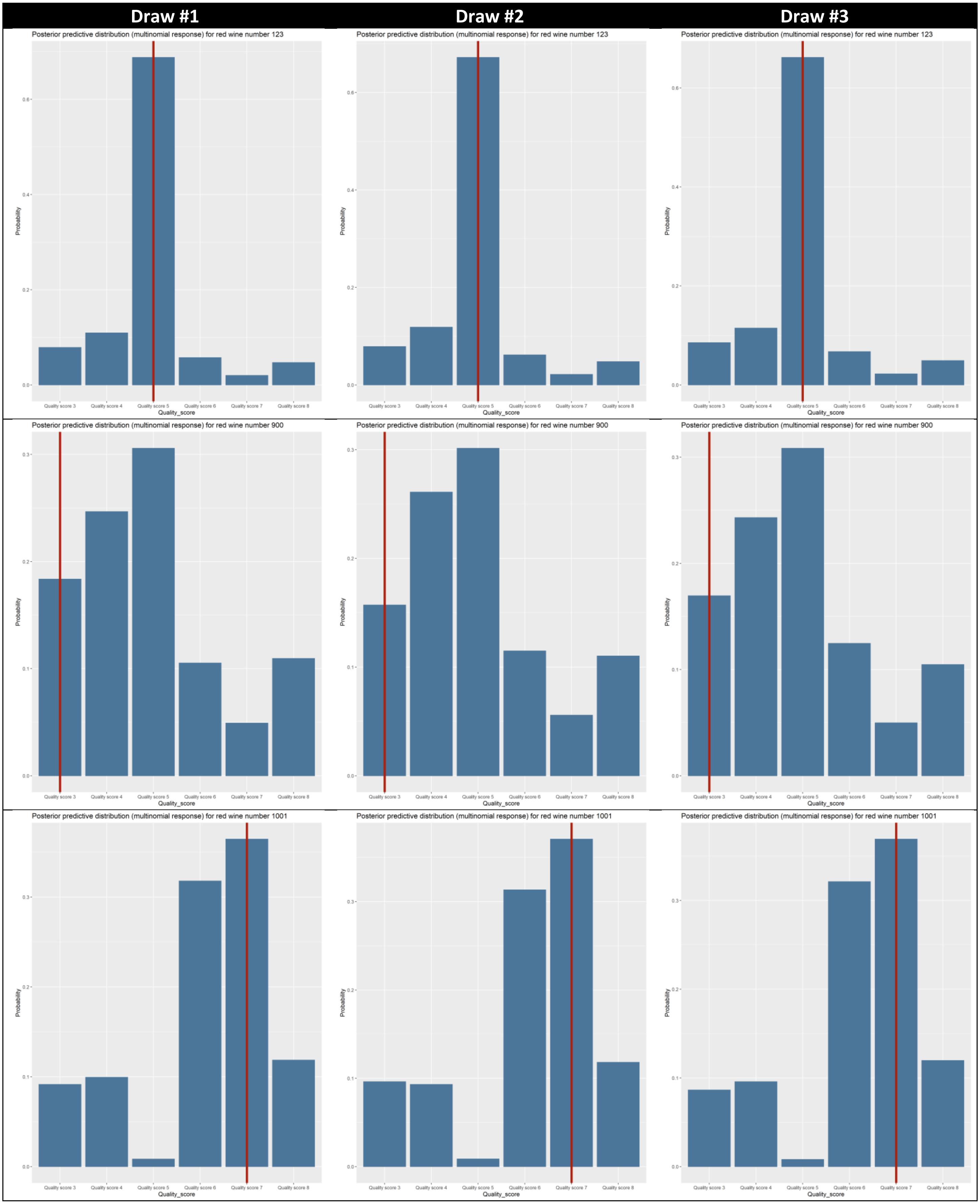
White - Multinomial. Posterior distribution of beta coefficients for Quality Score 9

White - Multinomial. Posterior distribution of beta coefficients for Quality Score 9

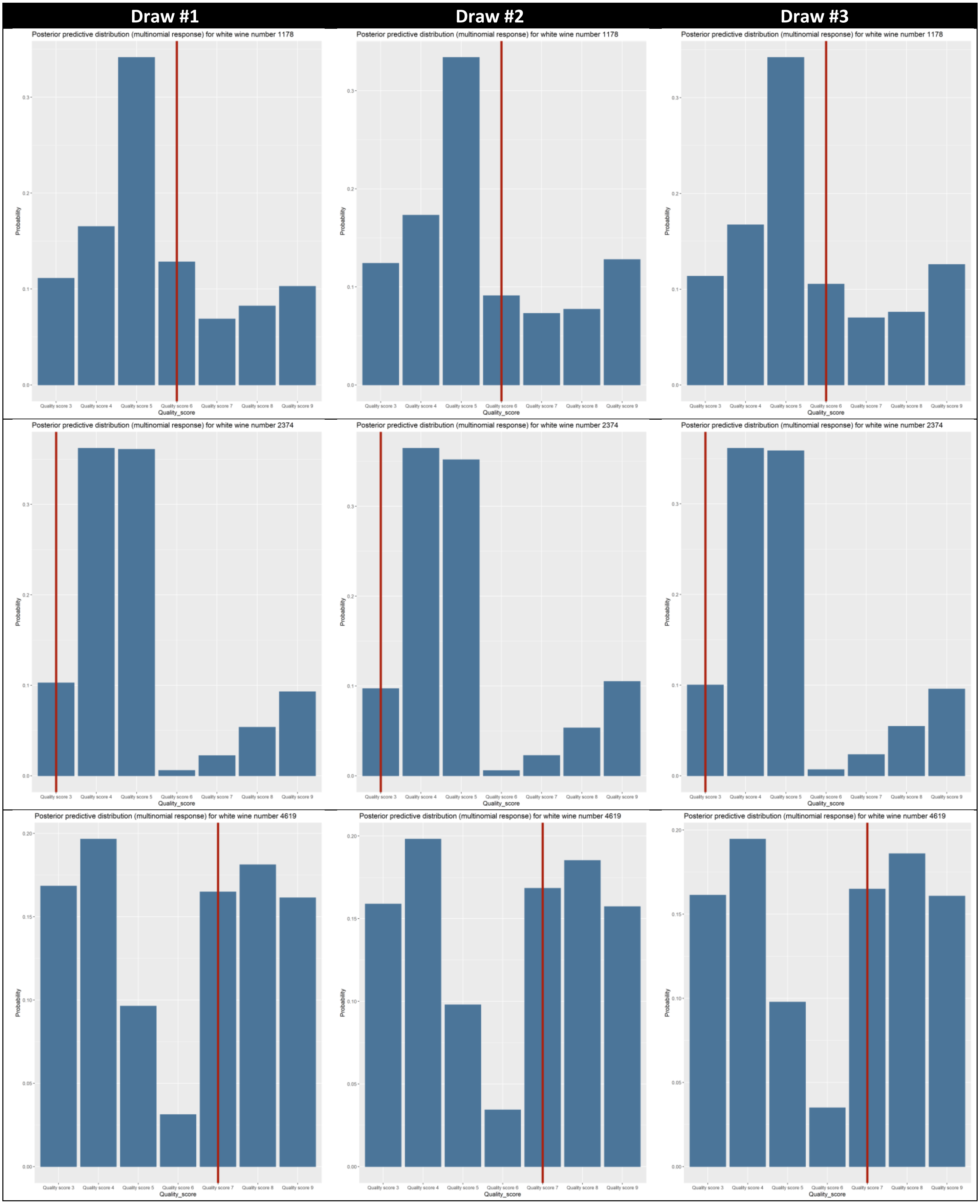
White - Multinomial. Posterior distribution of beta coefficients for Quality Score 9

APPENDIX 1. Posterior predictive distribution of Multinomial response model

Posterior predictive distribution of Multinomial response model – Red wine



Posterior predictive distribution of Multinomial response model – White wine


Draw #1
Draw #2
Draw #3
Posterior predictive distribution (multinomial response) for white wine number 1178
Posterior predictive distribution (multinomial response) for white wine number 2374
Posterior predictive distribution (multinomial response) for white wine number 4619
Probability
Quality score 3
Quality score 4
Quality score 5
Quality score 6
Quality score 7
Quality score 8
Quality score 9
Quality_score