INTRODUCTION

The models described in this report were created with the aim to decide which explanatory variables are relevant to establishing the quality of wine.

The data set is composed of 12 variables reporting the physicochemical properties of red and white Portuguese wine "Vinho Verde". The explanatory variables are:

- Fixed Acidity
- Volatile Acidity
- Citric Acid
- Residual Sugar
- Chlorides
- Free Sulphur Dioxide
- Total Sulphur Dioxide
- Density
- pH
- Sulphates
- Alcohol
- Quality

Quality is the response variable and represents a score assigned to a wine form a minimum of 0, which indicates a poor wine, to a maximum of 10 which indicates an excellent wine.

After a brief exploratory analysis which provides information about the general features of the dataset, there will be an analysis of different models containing all the explanatory variables. The models are created with the function *stan_glm* of the package *rstanarm*, in order to evaluate different parameters of the prior distribution for a better selection of the most appropriate one. After that, there will be variable selection and an analysis of the selected model. Initially, a Metropolis-Hasting sampler were created but after a strange behaviour in the output, has been decided to apply the function *blasso* from the package *monomvn*. The next step is to analyse a model created after applying a log transformation to all the variables of the dataset. It has also been tried to augment the number of variables by performing pairwise multiplication and ratio and analyse a model with all the variables included, and one after variable selection using lasso method. Finally, a comparison between all the models created has been performed using the functions *loo* and *loo_compare*, of the package *loo.* The procedure described above has been first applied to the red wine dataset and then to the white wine dataset.

EXPLORATORY ANALYSIS RED WINE

Table 1 reports all the statistical parameters of the explanatory and response variables and Figure 1 shows the histogram of red wine *quality* score. From the table it is possible to observe that the response variable goes from a minimum of 3 to a maximum of 8, with a mean of 5.6 and a median of 6. From the histogram it is instead possible to distinguish that the majority of wines have a score of 5 and 6.

	Fixed Acidity	Volatile Acidity	Citric Acid	Residual Sugar	Chlorides	Free Sulphur Dioxide
Min	4.60	0.12	0.00	0.90	0.01	1.00
1st Quartile	7.10	0.39	0.09	1.90	0.07	7.00
Median	7.90	0.52	0.26	2.20	0.08	14.00
Mean	8.32	0.53	0.27	2.54	0.09	15.87
3rd Quartile	9.20	0.64	0.42	2.60	0.09	21.00
Max	15.90	1.58	1.00	15.50	0.61	72.00
	Total Sulphur Dioxide	Density	pH	Sulphates	Alcohol	Quality
Min	6.00	0.9901	2.74	0.33	8.40	3.00
1st Quartile	22.00	0.9956	3.21	0.55	9.50	5.00
Median	38.00	0.9968	3.31	0.62	10.20	6.00
Mean	46.47	0.9967	3.31	0.66	10.42	5.636
3rd Quartile	62.00	0.9978	3.40	0.73	11.10	6.00
Max	289.00	1.0037	4.01	2.00	14.90	8.00

Table 1: *In the table are listed the minimum value, the first quartile, the median, the mean, the third quartile and the maximum value for the explanatory and response variables.*

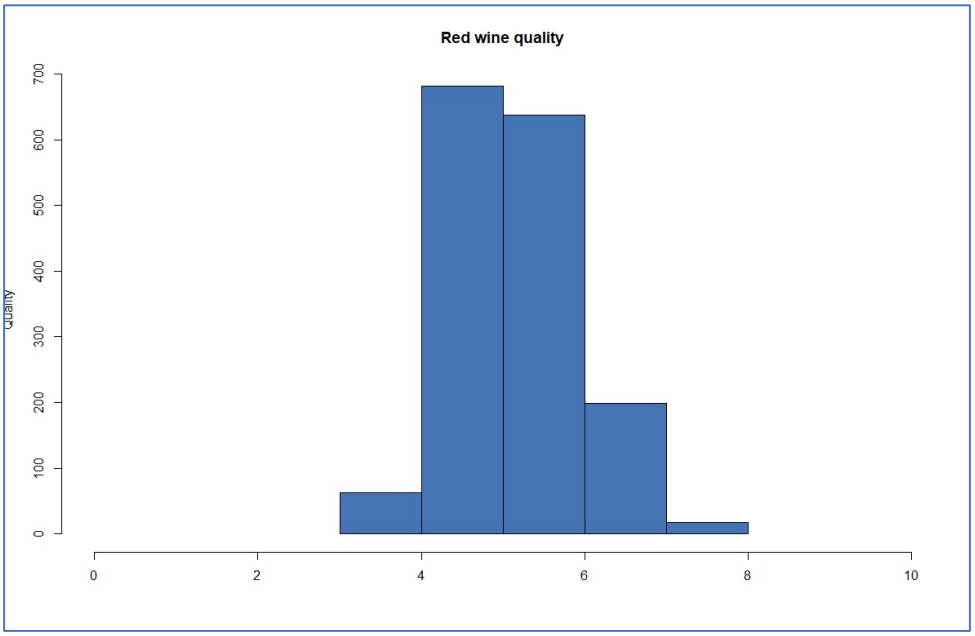

Figure 1: *Histogram for the red wine quality.*

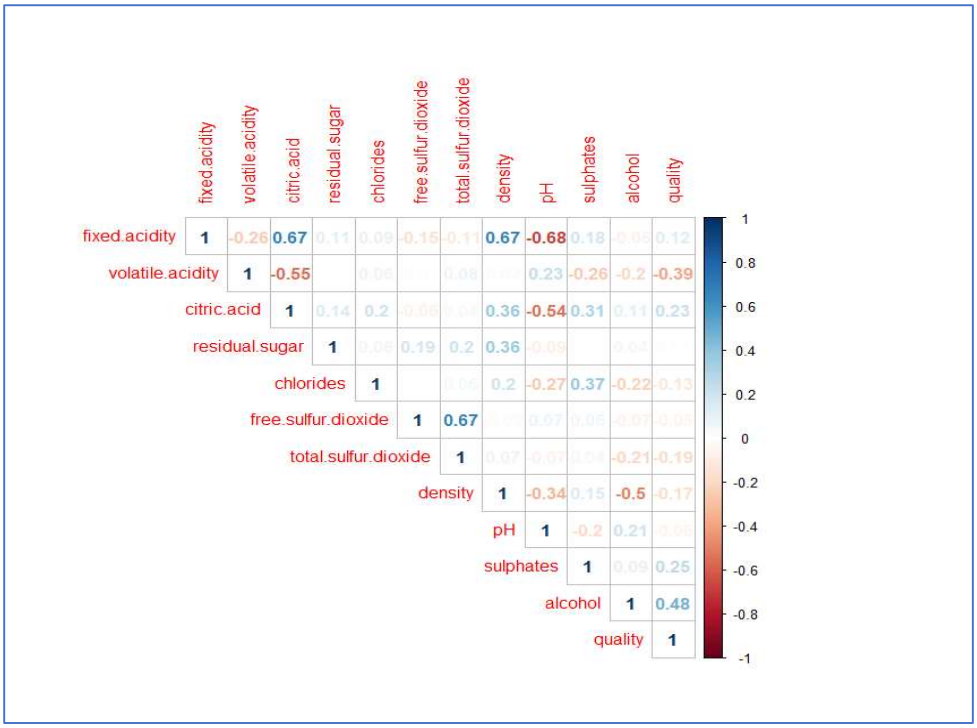

Figure 2: *Correlation matrix.*

Figure 2 shows the correlation matrix of the variables in the dataset. The highest positive correlation is 0.67 and it is between *density* and *fixed.acidity*. The same value is also between *total.sulfur.dioxide* and *free.sulfure.dioxide* and between *fixed.acidity* and *citric.acidity*. These last two correlations are obviously high because one of the variables involved is a measure of the other. Similarly, the high negative correlation of 0.68 of *pH* and *fixed.acidity* is due to the fact that *pH* is a measure of the acidity.

The response variable *quality* has not strong correlation with any of the variables. The highest positive correlation is 0.48 and is with *alcohol,* while the highest negative correlation is 0.39 and is with *volatile.acidity.*

ANALYSIS

The model describing the data that are object of the analysis in this report can be expressed as follow:

$$y = X\beta + \epsilon$$

Where:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{1599} \end{bmatrix},\ X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,11} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,11} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1599,1} & x_{1599,2} & \cdots & x_{1599,11} \end{bmatrix},\ \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{11} \end{bmatrix} \text{ and } \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{1599} \end{bmatrix}$$

The priors can be very influential in Bayesian analysis; therefore, it is a good idea to choose weakly informative priors unless there are good reasons available to establish that some parameters can come from the informative distribution specified by the prior. For this reason, the function *stan_glm* from the

package *rstanarm* has been used to test different prior distributions. The distribution that can be utilised is the normal distribution because the data can be well described by this distribution. For the intercept and the other regression parameters, different parameters of a normal distribution prior have been tested: 0 and 1 for the mean and 5, 10 and 15 for the standard deviation of the intercept; 0 and 1 for the mean and 2, 2.5 and 5 for the standard deviation of the other regression parameters. All the six models explored seem to be equivalent, hence, a model (called *model_6*) with the lowest values of $\hat{R}$ (see page 6 for an explanation of $\hat{R}$) has been chosen. This model has a prior distribution $N(1, 15)$ for the intercept and a prior distribution $N(1, 5)$ for the other regression parameters. These values for the priors will be used also for other models created for the red wine dataset.

The function *stan_glm* uses automatically adjusted priors to make sure that the specified priors are not too informative. The adjusted scales are calculated as follow:

$$Intercept = \sigma_i * sd(y)$$

$$Cofficients = \frac{\sigma_p}{sd(d)} * sd(y)$$

Where σ_i is the standard deviation of the prior for the intercept, σ_p is the standard deviation for the prior for the other regression parameters, $sd(y)$ is the standard deviation of the response variable and $sd(d)$ is the standard deviation of the data.
Figure 3 shows the traces of chains of this model and it is noticeable that for all the explanatory variables they are well mixed.

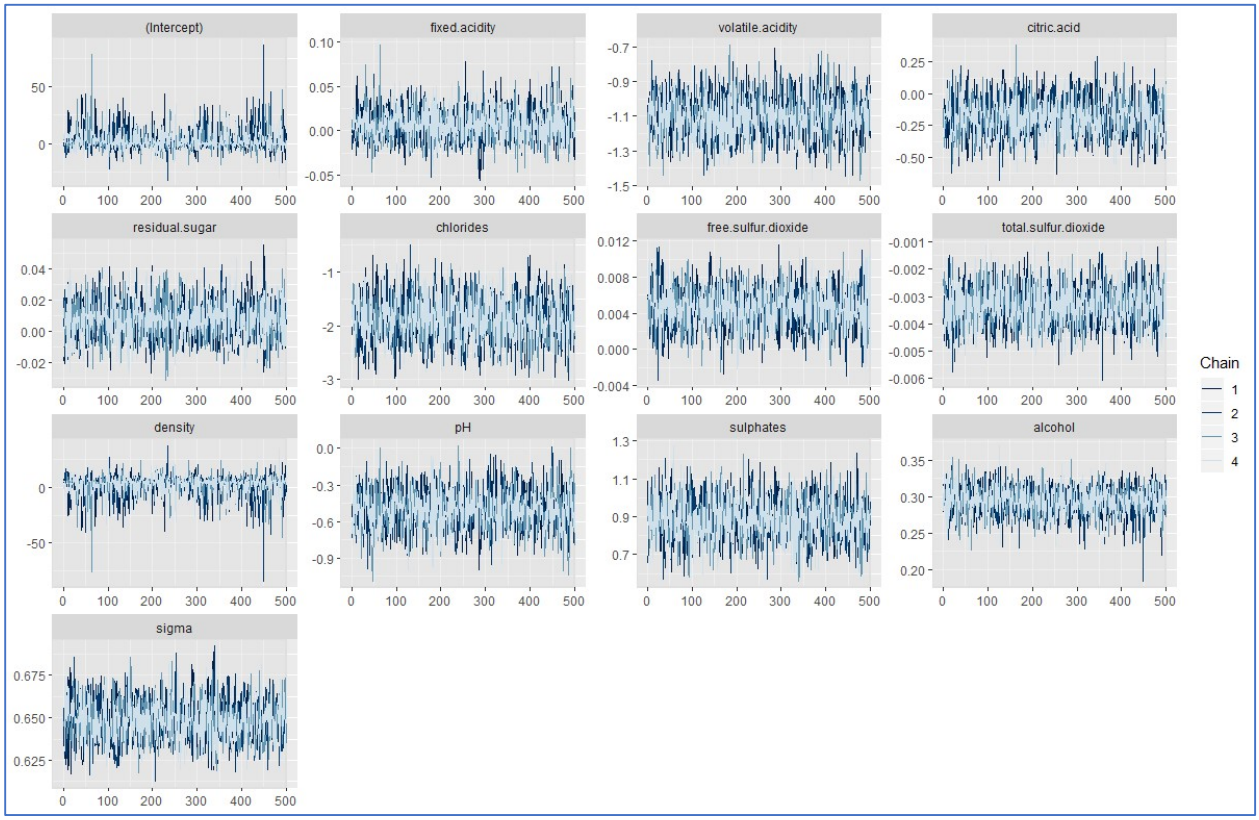

Figure 3: *Chain traces for all the variables of the dataset. The chains are well mixed.*

Furthermore, the Gelman-Rubin statistics can be used as additional proof of convergence of the models. $\hat{R}$ is basically a comparison between the variance within the chains and across the chains and it is calculated as follow:

$$\hat{R} = \sqrt{\frac{\widehat{var}(\theta)}{W}}$$

Where W is the mean of the variance of each chain. Figure 4 shows the $\hat{R}$ for the models created. They are very close to 1 and this also demonstrates that the 6 models are equivalent.

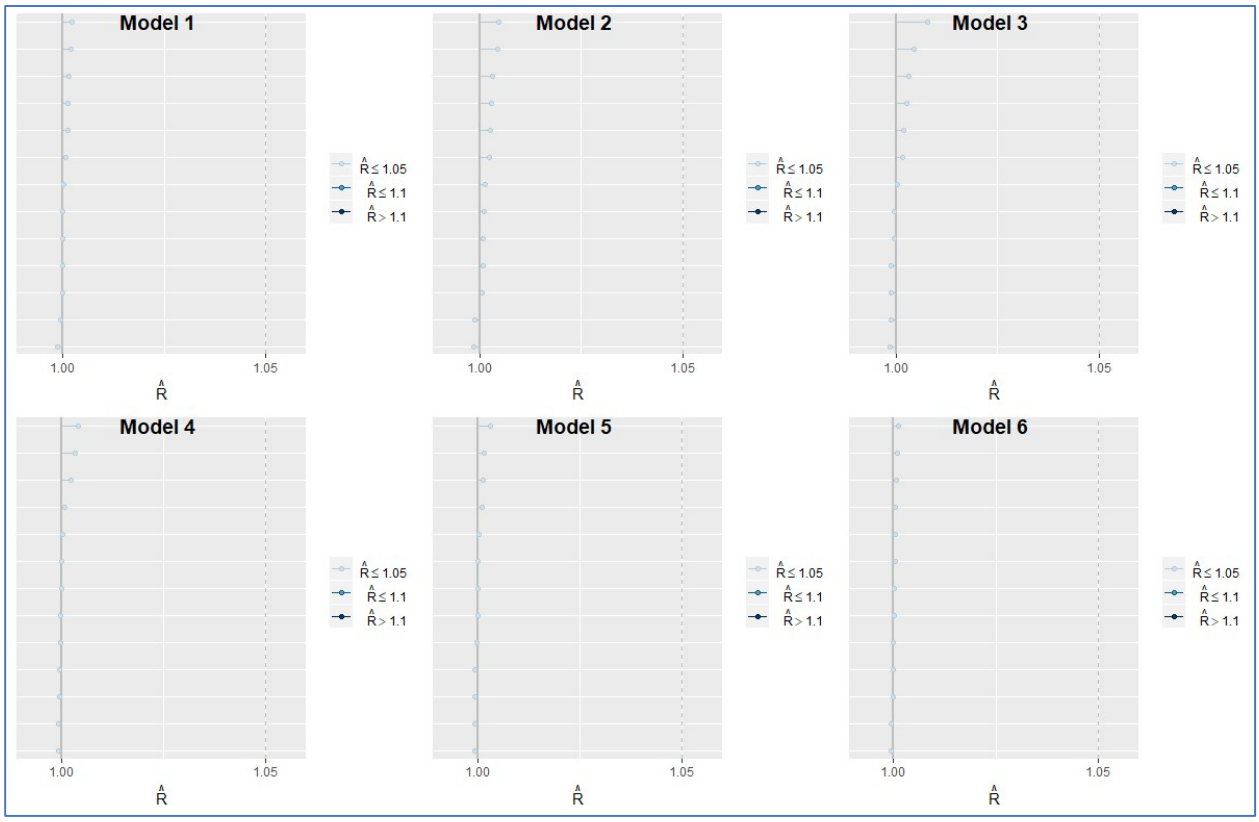

Figure 4: *$\hat{R}$ value for all the covariates in the datasets and for all the models created. The values are all near 1, therefore, the model has converged.*

Another important control to effectuate on the model chosen is the autocorrelation within each chain for all the variables. The samples are coming from a Markov Chain, therefore, they are correlated. However, figure 5 shows that, in the model analysed, there is no autocorrelation.

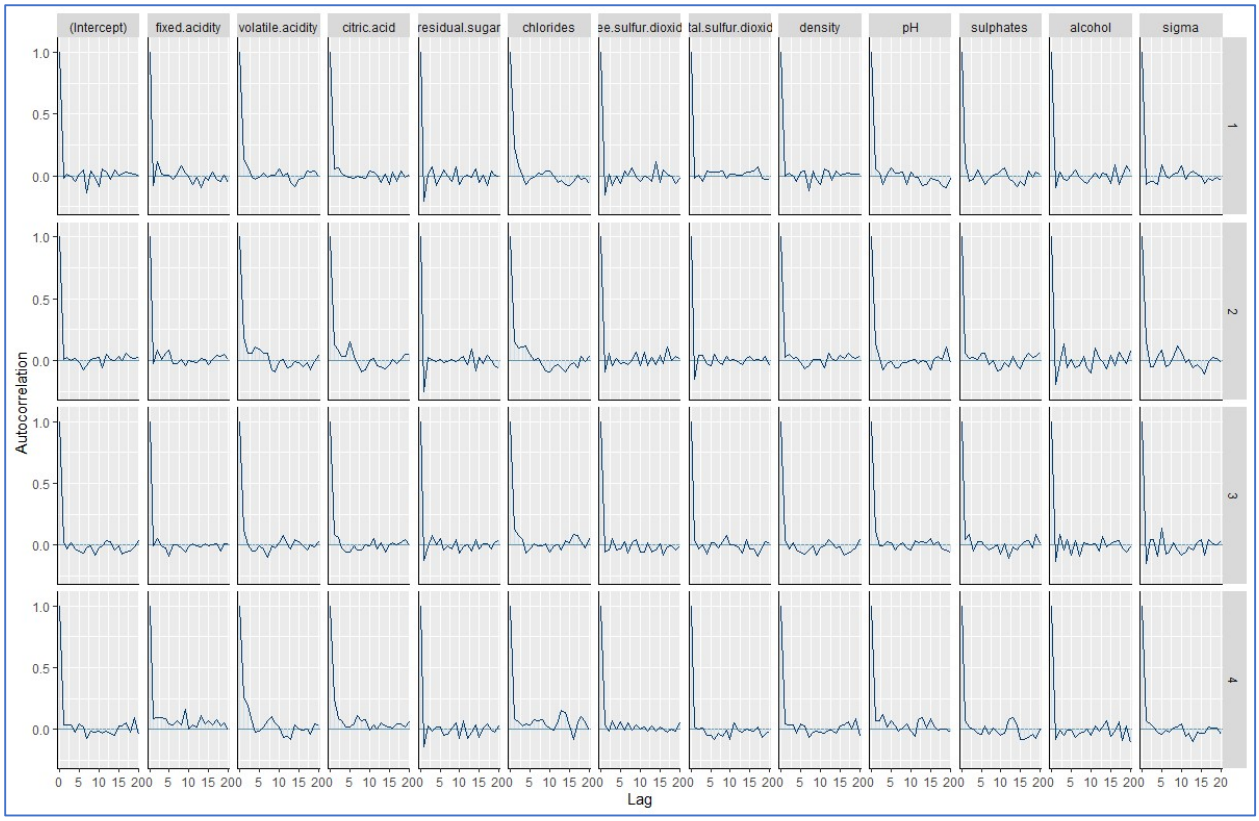

Figure 5 : *Autocorrelation for all the covariates in the dataset and for the four chains. They show no autocorrelation.*

VARIABLE SELECTION

After being sure that the model works well, variable selection can be performed, in order to control if all the covariates are significant for the response. A Bayesian lasso method using a Metropolis-Hasting sampler has been tried to be implemented. The hierarchical model describing the data is:

$$y \mid \beta_0, \beta, \sigma^2 \sim N(\beta_0 1_n + X\beta, \sigma^2 I)$$
$$\sigma^2 \sim Inv - Gamma(a, b)$$
$$\beta_0, \propto 1$$
$$\beta \mid \sigma^2, \tau_1^2, \dots, \tau_p^2 \sim N(0_k, \sigma^2 D_\tau)$$
$$\tau_1^2 \mid \lambda^2 \sim Exponential\ (0.5\lambda^2)$$
$$\lambda^2 \sim Gamma(a_\lambda, b_\lambda)$$

The data have been rescaled in order to have zero mean and unit variance. The hyperparameters for the prior on σ^2 have been chosen $a = 1, b = 15$, the same parameters of the model above chosen (*model_6)*. The hyper parameters on λ^2 have been selected to be $a_\lambda = 0.02,\ b_\lambda = 0.1$. The proposed new value is normally distributed and centered around the old value with a standard deviation of 0.1.

The central posterior intervals for the regression coefficients are shown in figure 6. As it is possible to notice the plot has a quite strange behavior and the estimated parameters might be not reliable. Therefore, the function *blasso* from the package *monomvn* has been applied, with the same values as before mentioned for the arguments *rd* and *ab.* The central posterior intervals are shown in figure 7.

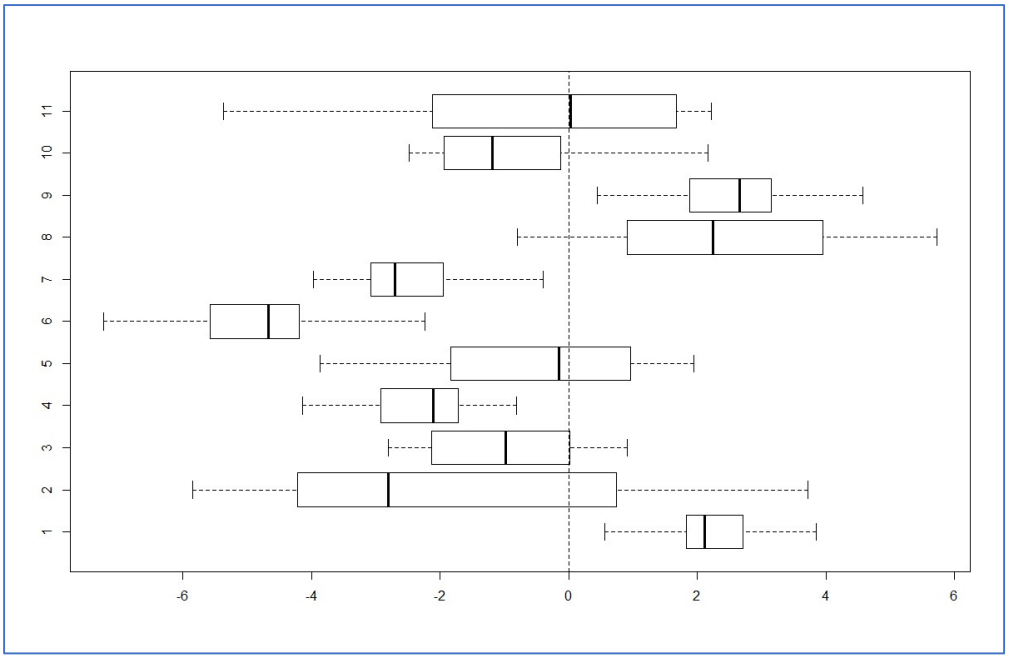

Figure 6: *Central posterior intervals for regression coefficients β using Metropolis-Hastings sampler.*

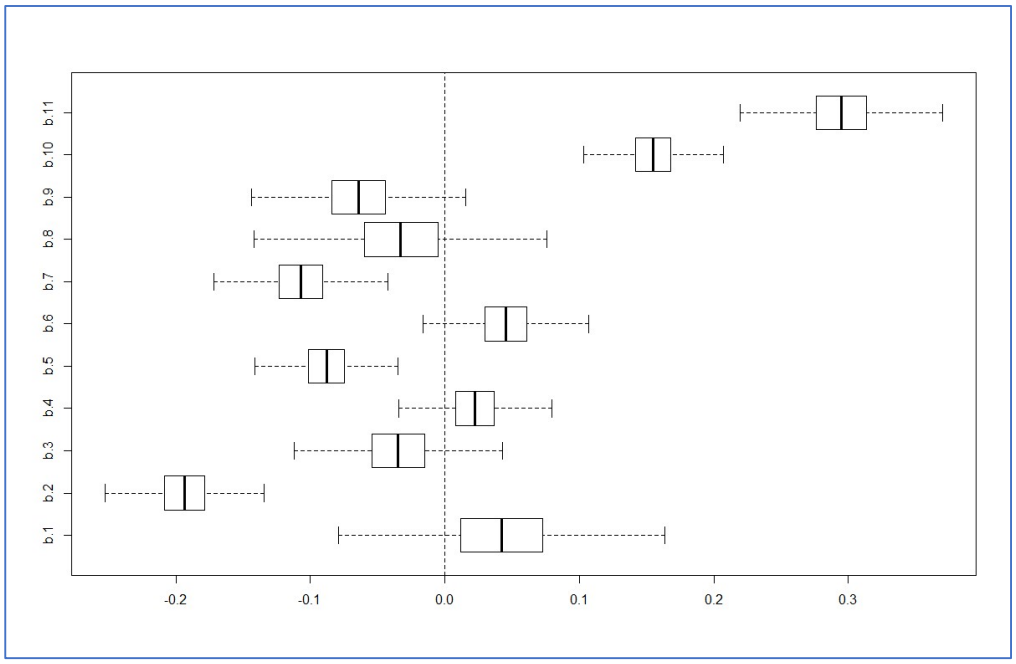

Figure 7: *Central posterior intervals for regression coefficients β using Gibbs sampler.*

The variables that can be excluded from the model are the ones that presents the estimate of the corresponding $\beta_i \mid y$ close to 0. Therefore, from figure 7, the closer to 0 are: $\beta_1, \beta_3, \beta_4$ and β_8, corresponding to the covariates: *fixed.acidity, citric.acid, residual.sugar* and *density*. A model with variables selected has been created, maintaining the same value for the parameters of the priors. The model has converged because the traces of the four chains have the same behaviour of figure 3 and the $\hat{R}$ values are shown in table 2. Besides, there are no signs of autocorrelation.

	Intercept	Volatile Acidity	Chlorides	Free Sulphur Dioxide	Total Sulphur Dioxide	pH	Sulphates	Alcohol	Sigma
$\hat{R}$	1.0003594	0.9997785	1.0047583	0.9988930	1.0036910	1.0015530	1.0061947	0.9998175	1.0001632

Table 2: *Values of $\hat{R}$ for all the covariates chosen. The values are all very close to 1, indicating that the chains have converged.*

PREDICTIONS

It is now possible to perform predictions using both models described above (*model_6* and *model_selected*), using the function *posterior_linpred* and *posterior_predict* from the package *rstanarm*. Both functions use posterior distributions to make predictions. The posterior distribution is composed of *n* number of iterations and it is possible to use posterior draws to calculate predicted *quality* of wine. For example, the parameters values at iteration 1 can be used to calculate predicted *quality* of wine, another set of predicted *quality* of wine can be obtained using the parameters values at iteration 2, and so on. At the end, *n quality* of wine for each wine can be calculated because *n* iterations are available.

The function *posterior_linpred,* returns a matrix which has as rows all the iterations and as columns all the observations. One check that can be performed is to compare the summary of the original data with the summary of the iterations. If the statistical parameters are similar the model can predict quite well the *quality* of wine.

The first model used for prediction is *model_6*, the one chosen between the first six models created for choosing the prior parameters, and it contains all the covariates of the data set. Table 3 shows this comparison between the original data of *quality* of wine and some of the iterations.

	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Original Data	3.000	5.000	6.000	5.636	6.000	8.000
Iteration n. 1	4.167	5.227	5.545	5.595	5.951	7.181
Iteration n. 328	4.191	5.230	5.560	5.621	5.990	7.358
Iteration n.500	4.183	5.266	5.602	5.646	6.013	7.489
Iteration n. 1025	4.093	5.265	5.595	5.630	5.978	7.289

Table 3: *Summary statistics for the response variable and some of the iterations obtained. The mean has similar values for all 5 data taken into consideration, whilst the range of Min and Max is shorter for the iterations. The Median has lower values for iterations.*

It is quite clear that the iterations have smaller range of variation. The *quality* of wine for the original data goes from a minimum of 3 to a maximum of 8. Instead, for the four iterations go from a minimum of 4.1 to a maximum of 7.5. Same situation for the median: for the original data is 6, whilst for the iterations is between 5.54 and 5.60. Only the mean has similar values for all five data taken into consideration.

Using the function *posterior_predict* it is possible to plot the posterior distribution for each wine and compare it with the real value of *quality*. In figure 8 are reported plots of the posterior distribution for wines that have a score of *quality* from 3 to 8. As it is possible to observe, the model overestimates the score of wine which originally had a lower score, and underestimates those with a higher score. It performs better for wines with medium scores of *quality.*

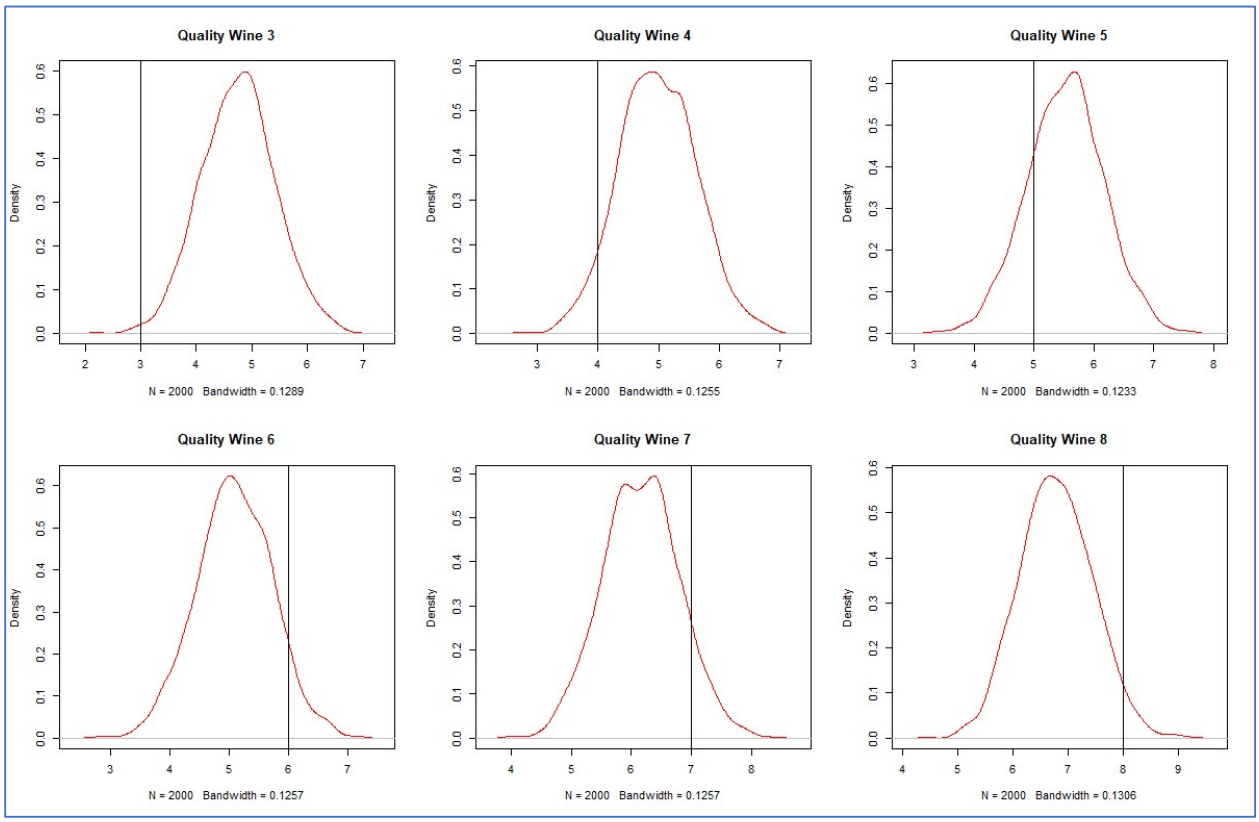

Figure 8 : *Posterior distributions for all the six class of quality of wine. The model overestimates quality of wine with lower scores and underestimates the ones with higher scores.*

One last step is to summarise results and make a global evaluation of how well the model fits the data. The plot in figure 9 shows all the posterior distributions and the observed data. It is possible to see that the distribution of the observed data (dark blue line) does not really fit all the other distributions (light blue lines). As it has been seen on the previous figure, this reveals that the model does not predict very well the quality of wine.

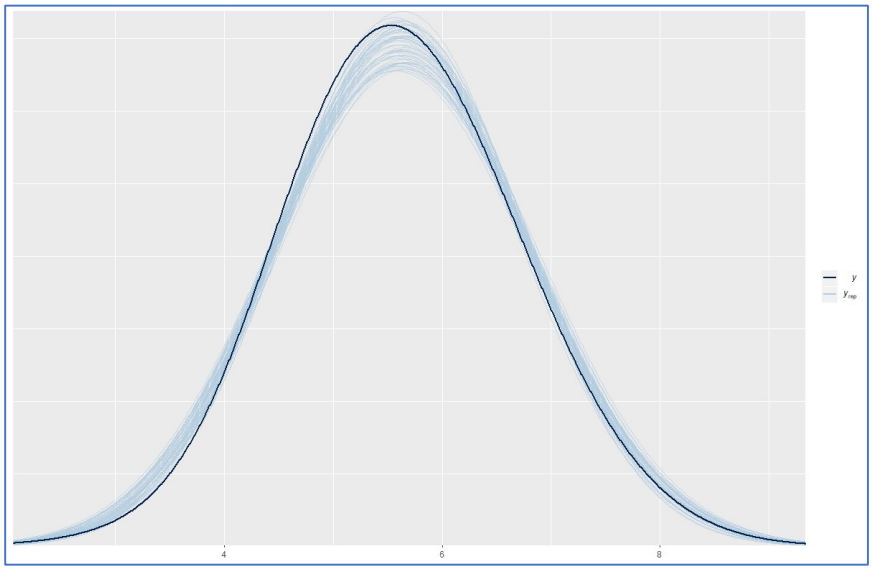

Figure 9 : *Distribution of observed data (dark blue line) does not properly fit the distribution of all the posteriors (light blue lines). That means that the model does not predict very well the quality of wine. Model containing all the covariates.*

The procedure above described has been also applied to the model (*model_selected*) created after the variable selection using the lasso method. As it is shown in figure 10, the model has the same behaviour of the previous one and from the posterior predictive distribution (not in this case reported) it is clear that the model overestimates the wine with lower score of *quality* and underestimates those with higher scores.

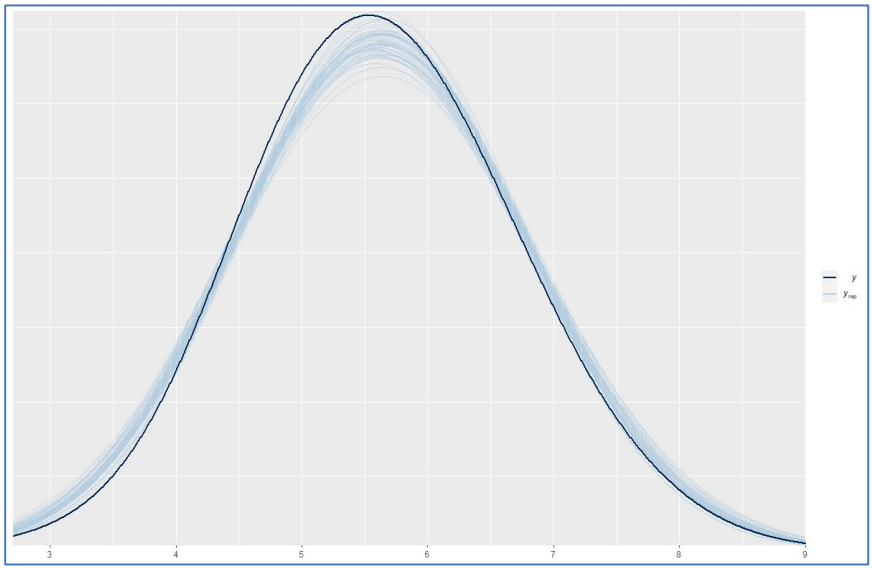

Figure 10: *Model created after the variable selection with the lasso method. Also for this model, the distribution of observed data (dark blue line) does not properly fit the distribution of all the posteriors (light blue lines). That means that the model does not predict very well the quality of wine.*

LOGARITHMIC TRANSFORMED MODEL

The data used for the model described above have been logarithmic transformed and a new model has been created. Besides, before the logarithmic transformation, a small amount (0.1) has been added to all the values that were equal to zero in order to avoid of obtaining NaNs (Not a Number) in the dataset. The four chains are well mixed and the values of $\hat{R}$ are near to 1 for all the covariates, meaning that the model has converged. There are also no signs of autocorrelation.

	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Original Data	1.099	1.609	1.792	1.719	1.792	2.079
Iteration n. 1	1.524	1.655	1.708	1.719	1.778	1.960
Iteration n. 328	1.511	1.660	1.713	1.723	1.779	1.996
Iteration n.500	1.509	1.650	1.708	1.720	1.783	2.012
Iteration n. 1025	1.498	1.646	1.707	1.716	1.780	1.995

Table 4: *Summary statistics for the response variable and some of the iterations obtained. The mean has similar values for all 5 data taken into consideration, whilst the range of Min and Max is shorter for the iterations. The Median has lower values for iterations.*

Table 4 shows the summary statistics for the logarithmic transformed model and it presents the same situation of the previous two models. For the observed data the values go from a minimum if 1.099 to a maximum of 2.079 whilst for the iterations they go from a minimum of 1.498 to a maximum of 2.012. Iterations have lower median than the observed data and only the mean has similar values for all 5 data taken into consideration.

Also the plots in figure 11 display that the model overestimates the *quality* of wine for those with lower scores and underestimates the ones with higher scores.

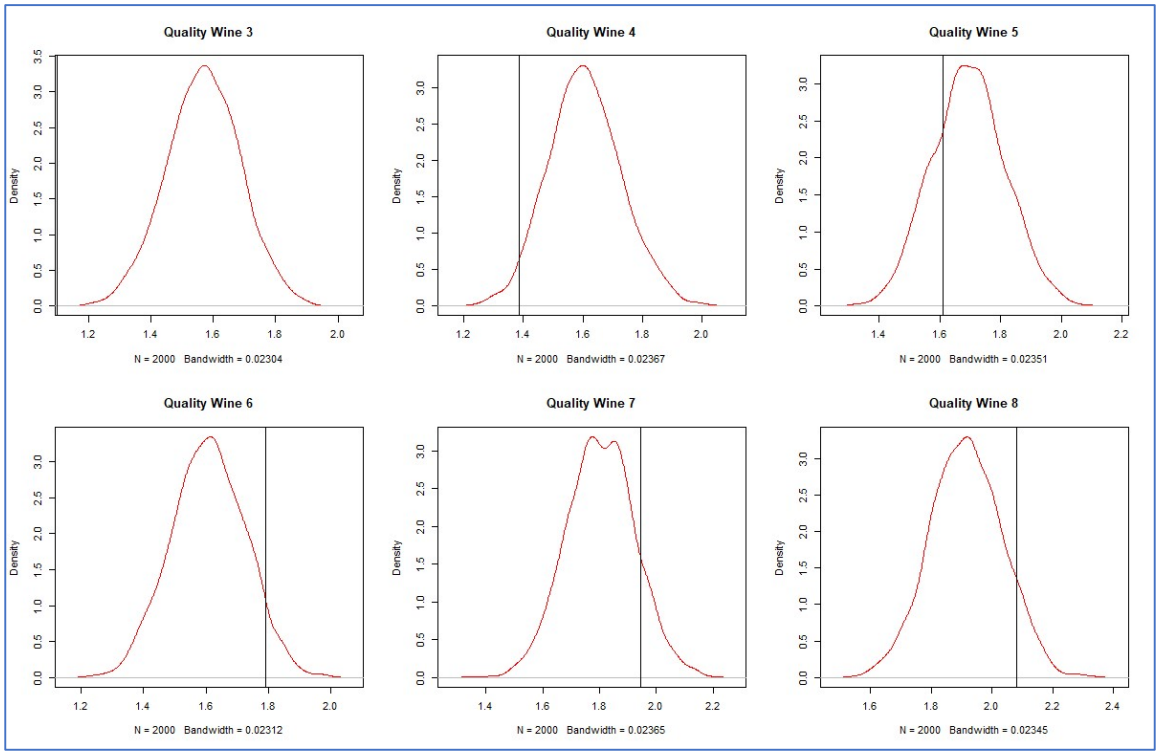

Figure 8 : *Posterior distributions for all the six class of quality of wine. The model overestimates quality of wine with lower scores and underestimates the ones with higher scores.*

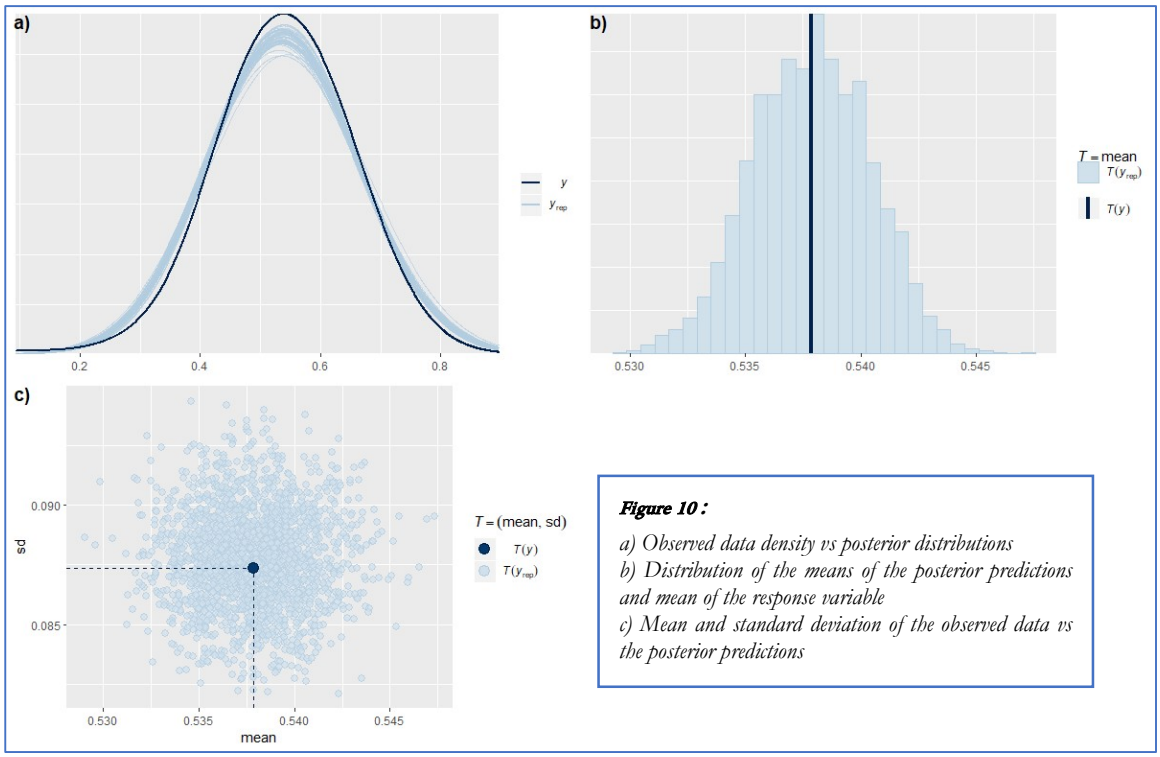

Figure 10 :

a) Observed data density vs posterior distributions
b) Distribution of the means of the posterior predictions and mean of the response variable
c) Mean and standard deviation of the observed data vs the posterior predictions

Figure 10 a) shows the density of the observed data vs the posterior predictions. In contrast with the plot of the two previous model, the mean seems to be more in accordance between the predicted posterior densities and the observed data. However, also for this model, the observed distribution does not perfectly fit the predicted ones. Figure 10 b) shows the distribution of the means of the posterior predictions (light blue bins) with superimposed the mean of the observed data (dark blue line) which it clearly lays in the middle of the expected range. Figure 10 c) shows the standard deviation vs the mean for the posterior predictions (light blue dots) and the observed data (dark blue dot). it is easily noticeable that the dark blue dot lays in the middle of the light blue ones. Even if it is not an excellent model for the reasons explained above, all the graphs form figure 10 indicate that the model may be a more effective choice for the data than the previous two models.

DATASET AUGMENTED WITH PAIRWISE MULTIPLICATIONS AND RATIOS

The original dataset has been augmented with pairwise multiplications and ratios. The covariates chosen for multiplications are: *citric.acid* and *total.slphur.dioxid* with a new covariate called *catsd; residua.sugar* and *pH* with a new covariate called *rspH; sulphates* and *alcohol* with a new covariate called *sa.* The covariates chosen for the ratios are: *fixed.acididty* and *chlorides* with a new covariate called *fac; density* and *pH* with a new covariate called *dpH; free.sulphure.dioxide* and *alcohol* with a new covariate called *fsda.*

The model has been implemented considering again the same values of the previous models for the parameters of the priors: for the intercept is $N(1, 15)$ and for the other regression parameters is $N(1, 5)$.

Figure 11 shows that the chains are well mixed and figure 12 shows that all the values of $\hat{R}$ for all the variables are near to 1, indicating that the model has converged.

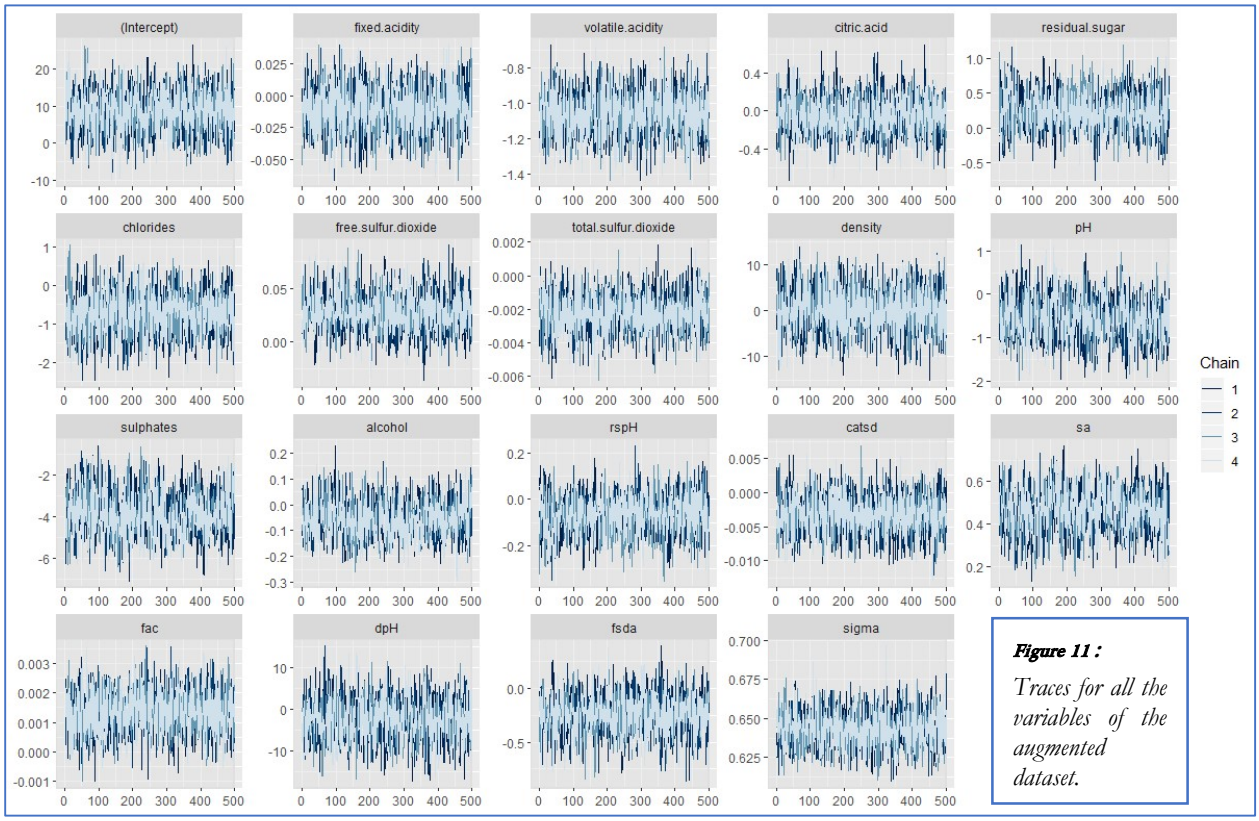

Figure 11 :
Traces for all the variables of the augmented dataset.

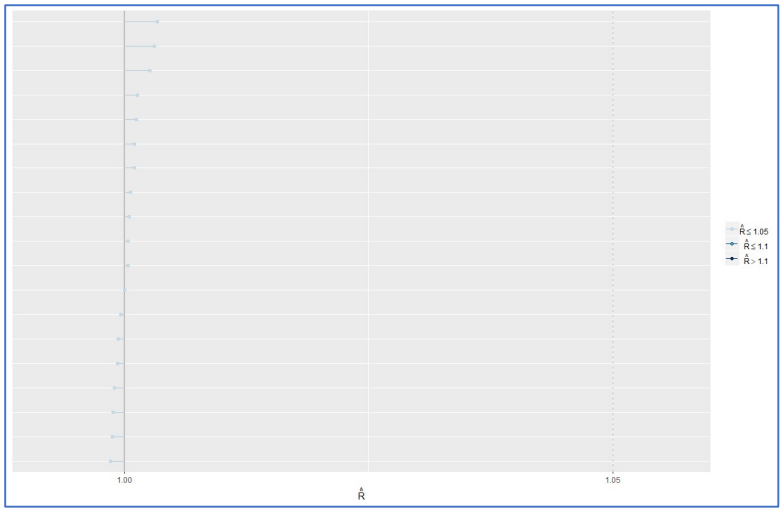

Figure 12 : *$\hat{R}$ for all the variables of the augmented dataset. All the values are very close to 1 indicating that the model has converged.*

Predictions with this model are characterised by the same feature of the previous models analysed. As it is possible to observe in figure 13, the model overestimates *quality* of wine for those with lower scores and underestimates the ones with higher scores.

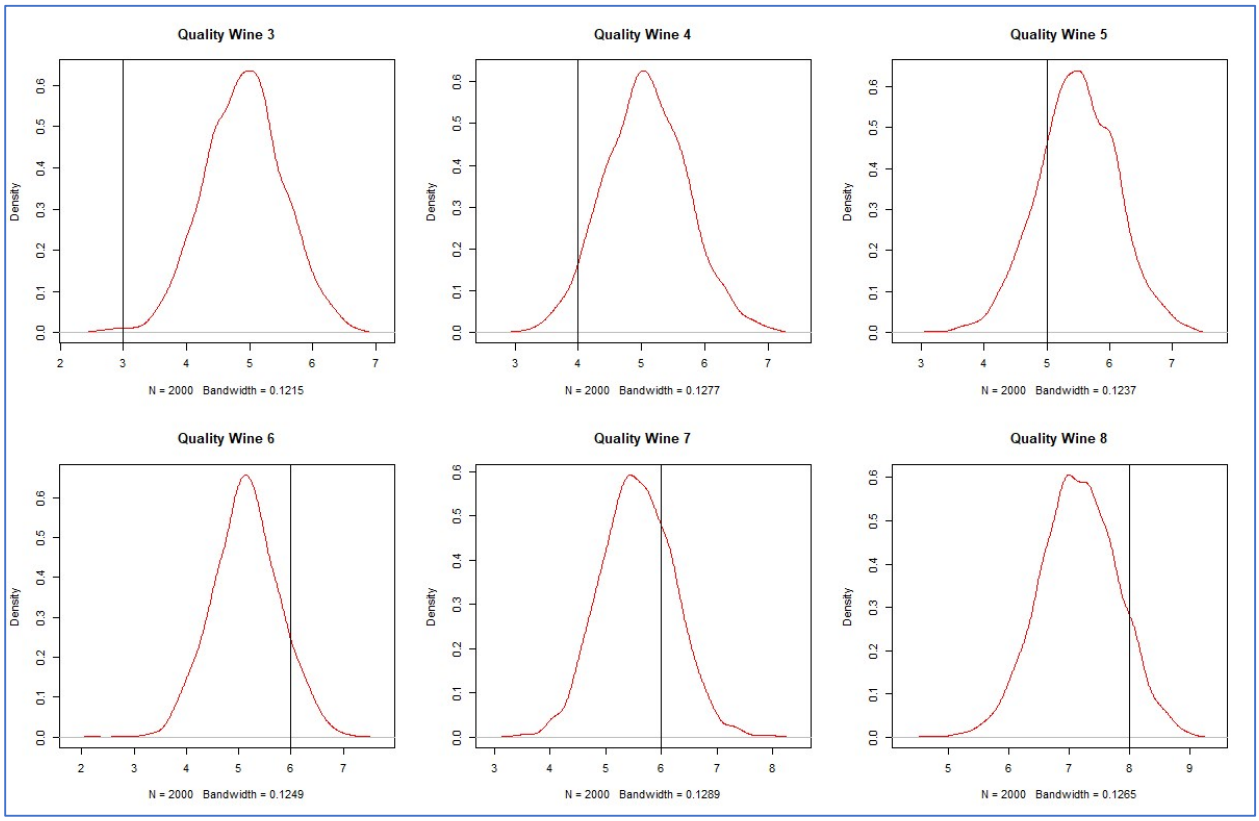

Figure 13 : *Posterior distributions for all the six class of quality of wine. The model overestimates quality of wine with lower scores and underestimates the ones with higher scores.*

Figure 14 illustrates that the observed data density does not fit properly the density of the posterior predictions, indicating that the model does not predicts very well the *quality* of wine.

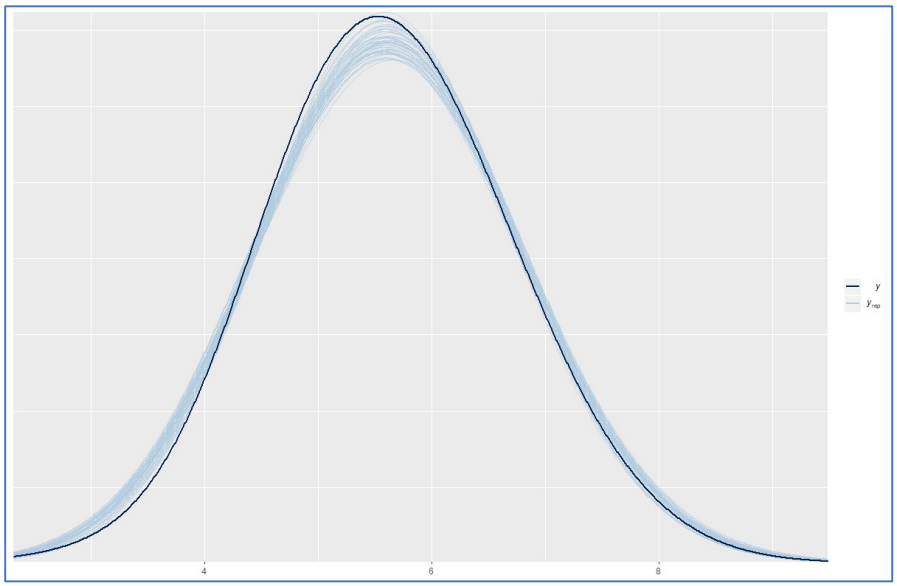

Figure 14: *Model created after adding pairwise multiplications and ratios. The figure shows that the distribution of observed data (dark blue line) does not properly fit the distribution of all the posteriors (light blue lines). That means that the model does not predict very well the quality of wine.*

At this point, the Bayesian lasso method using the Gibbs sampler through the function *blasso* of the package *monomvn,* has been reapplied to check if all the variables added with pairwise multiplications and ratios are effectively relevant to predict the *quality* of wine. In figure 15 are reported all the central posterior intervals obtained. The covariates that are close to 0, will be removed from the dataset and a last model will be analysed.

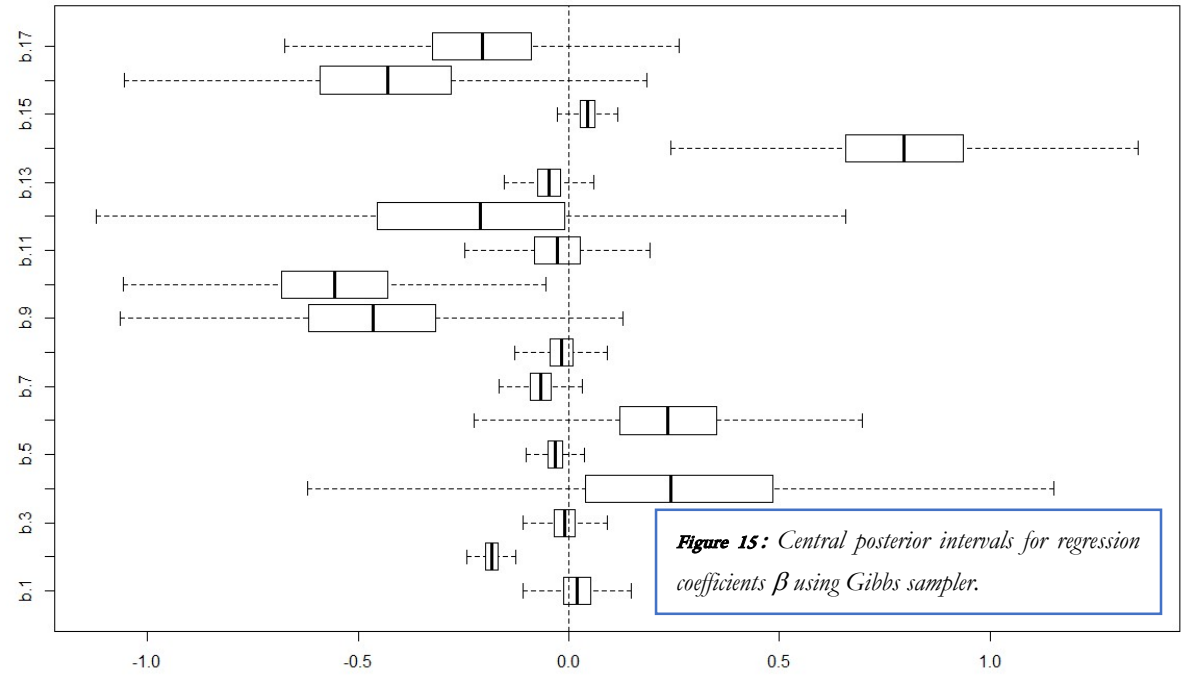

Figure 15: *Central posterior intervals for regression coefficients β using Gibbs sampler.*

Form figure 15 it is clear that the variable that can be dropped from the dataset are $\beta_1, \beta_3, \beta_8$ and β_{11} corresponding to *fixed.acidity, citric.acid, density* and *alcohol.*

The model presents the traces of the four chains well mixed and $\hat{R}$ values very close to 1 indicating that the model has converged. Analysing for this model the same type of plots seen so far, it is evident that the same problems reported for previous models are also here present. The model overestimates *quality* of wine with lower scores and underestimates those with higher scores.

WHITE WINE DATASET

Table 5 reports all the statistical parameters of the explanatory and response variables and Figure 16 shows the histogram of white wine quality score. From the table it is possible to observe that the response variable goes from a minimum of 3 to a maximum of 9, with a mean of 5.88 and a median of 6. From the histogram it is instead possible to distinguish that the majority of wines have a score of 6.

	Fixed Acidity	Volatile Acidity	Citric Acid	Residual Sugar	Chlorides	Free Sulphur Dioxide
Min	3.800	0.800	0.0000	0.600	0.00900	2.00
1st Quartile	6.300	0.2100	0.2700	1.700	0.03600	23.00
Median	6.800	0.2600	0.3200	5.200	0.04300	34.00
Mean	6.855	0.2782	0.3342	6.391	0.04577	35.31
3rd Quartile	7.300	0.3200	0.3900	9.900	0.05000	46.00
Max	14.200	1.1000	1.6600	65.800	0.34600	289.00
	Total Sulphur Dioxide	Density	pH	Sulphates	Alcohol	Quality
Min	9.0	0.9871	2.720	0.2200	8.00	3.00
1st Quartile	108.0	0.9917	3.090	0.4100	9.50	5.00
Median	134.0	0.9937	3.180	0.4700	10.40	6.00
Mean	138.0	0.9940	3.188	0.4898	10.51	5.878
3rd Quartile	167.0	0.9961	3.280	0.5500	11.40	6.00
Max	440.0	1.0390	3.820	1.0800	14.20	9.00

Table 5: *In the table are listed the minimum value, the first quartile, the median, the mean, the third quartile and the maximum value for the explanatory and response variables.*

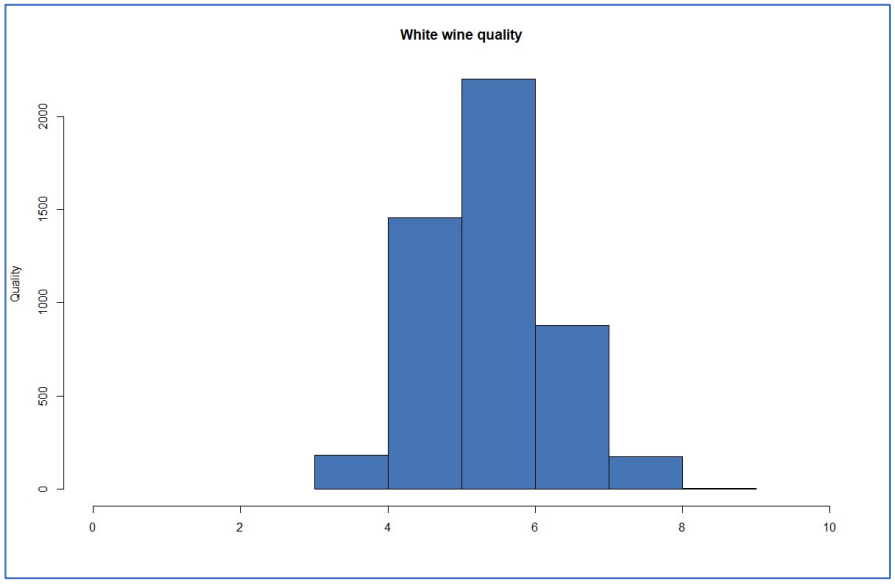

Figure 16: *Histogram for the white wine quality.*

The model describing this dataset is the same for the red wine dataset (see page 4). The function *stan_glm* of the package *rstanarm* has been applied to create, also in this case, 6 models with different parameters of a normal prior distributions. The normal distribution is considered the best in describing also this dataset.

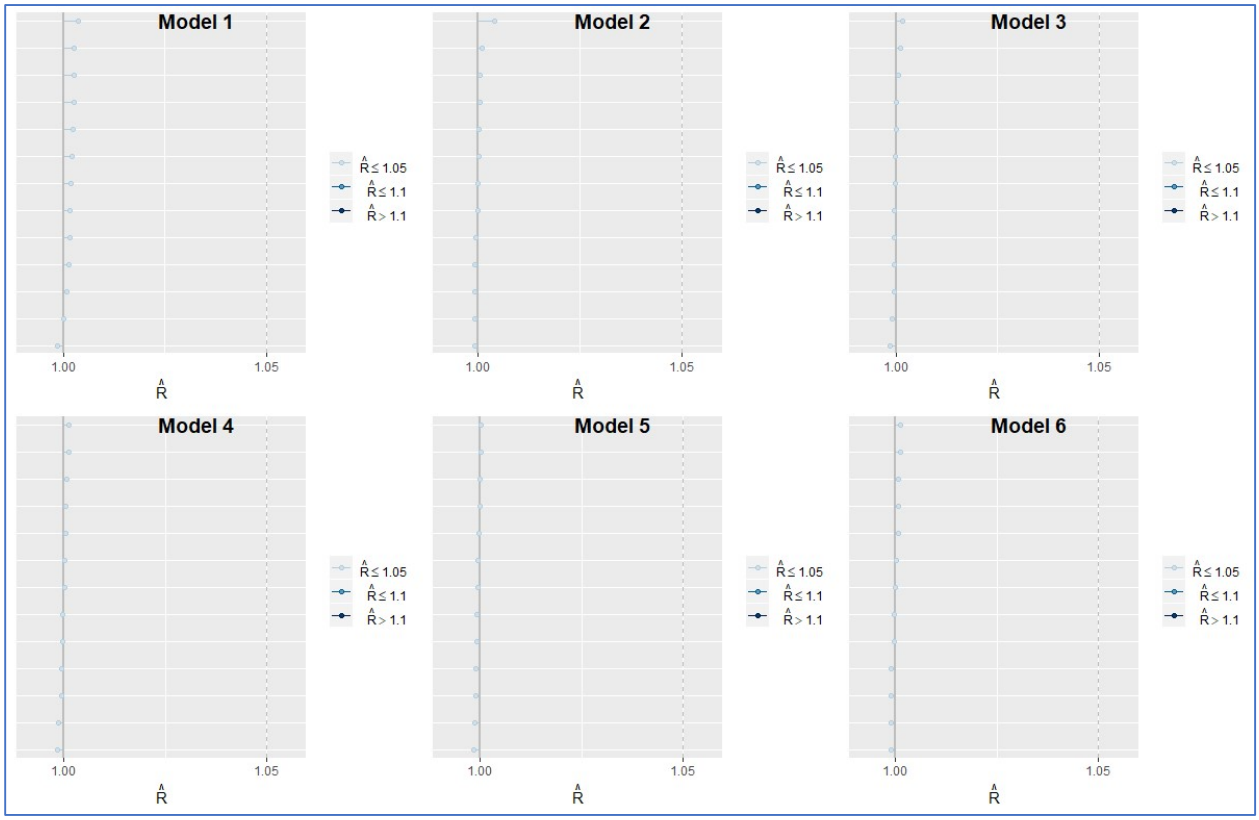

Figure 17: $\hat{R}$ *value for all the covariates in the datasets and for all the models created. The values are all near 1, therefore, the model has converged.*

Figure 17 shows that also in this case the six models seem to be equivalent, therefore, the model with the lowest values of $\hat{R}$ has been chosen. *model_3* has a prior distribution $N(1,5)$ for the intercept and a prior distribution $N(1,2)$ for the other regression parameters. Predictions using *model_3* present the same situation of red wine dataset: the model overestimates the *quality* of wine with lower scores and underestimate the ones with higher scores.

The function *blasso* from the package *monomvn* has been applied, with the same values as before mentioned for the arguments *rd* and *ab* and the central posterior intervals are shown in figure 18. The variable that can be dropped from the dataset are β_3, β_5 and β_7 corresponding to the covariates *citric.acid, chlorides* and *total.solphure.dioxide.* Another model (*model_select*) has been created without considering the covariates above listed.

From figure 19 it is possible to notice that for both models (*model_3* and *model_select*) the density distribution for the observed data do not fit the posterior predictive distribution densities, therefore, both models do not predict very well the *quality* of wine.

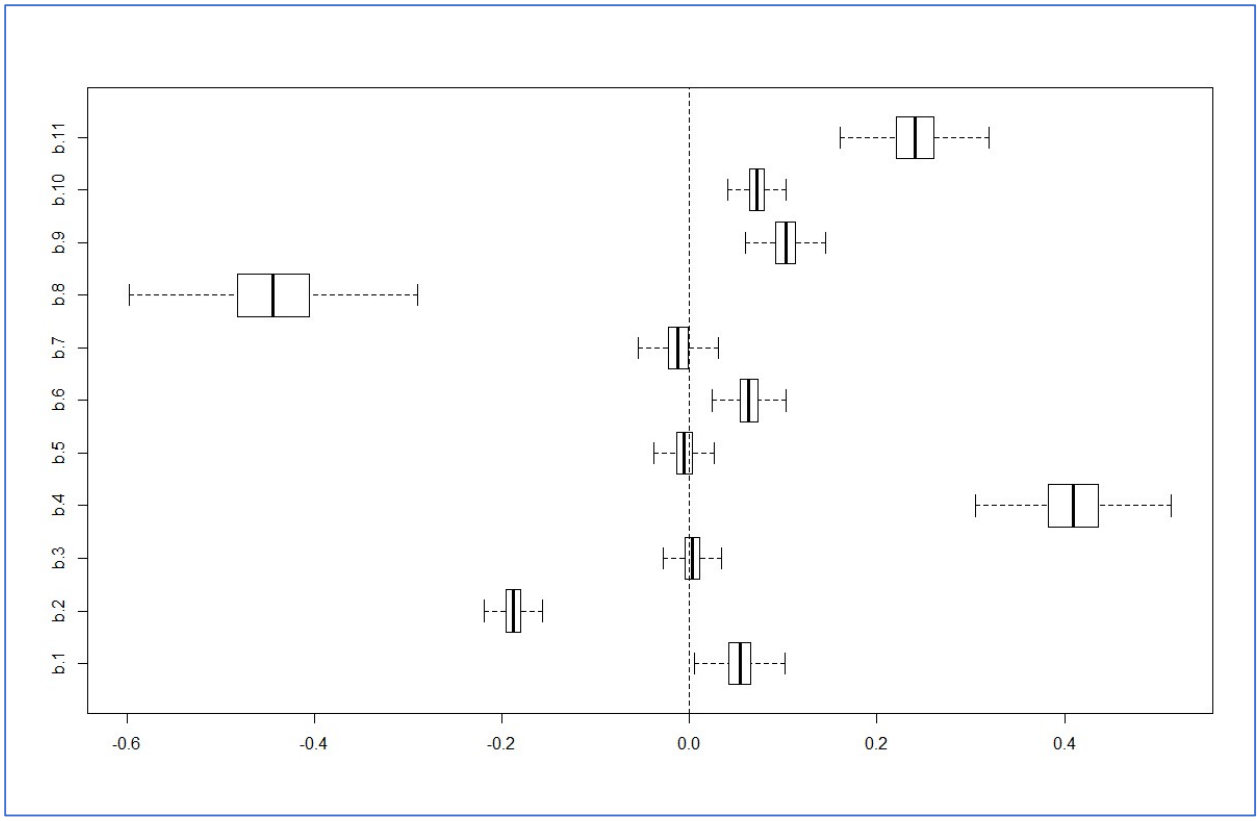

Figure 18 : *Central posterior intervals for regression coefficients β using Gibbs sampler.*

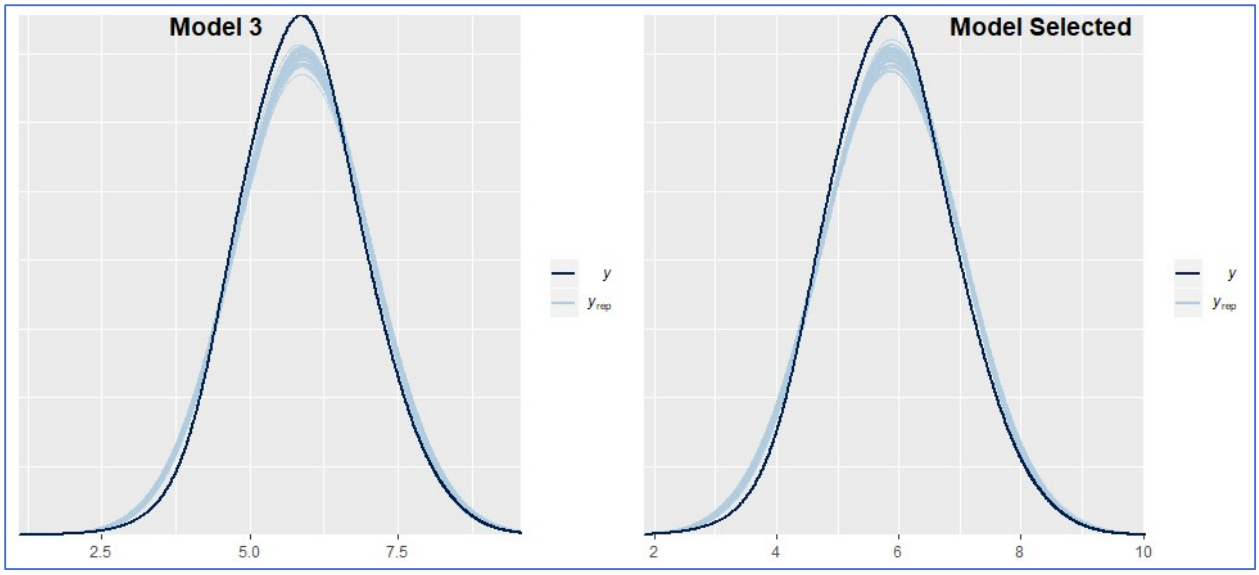

Figure 19 : *Distribution of observed data (dark blue line) does not properly fit the distribution of all the posteriors (light blue lines) for Model_3 and Model_select. This means that both models do not predict very well the quality of wine.*

LOGARITHMIC TRANSFORMED MODEL

The original data for white wine has been logarithmic transformed and the predictions obtained from this model are shown in figure 20. It seems that this model slightly better predicts the *quality* of wine. It is still overestimating the wine with lower scores but it also seems doing a better job with wines with higher scores.

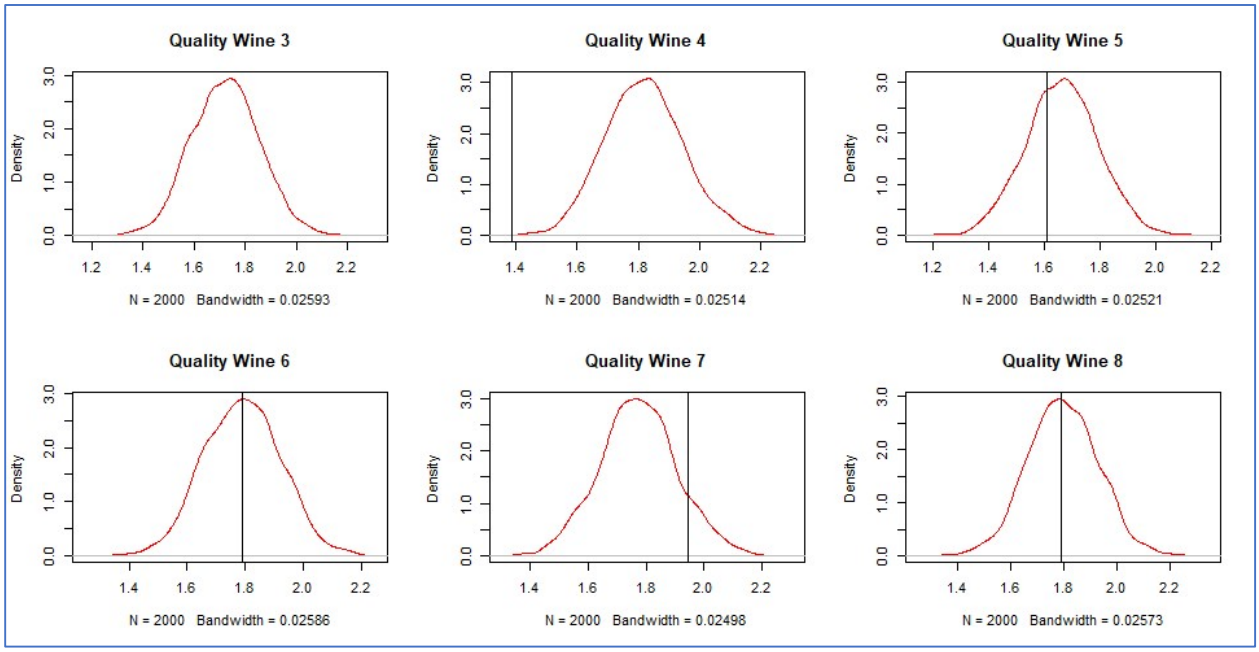

Figure 20 : *Posterior distributions for all the six class of quality of wine. The model overestimates wine quality with low scores and slightly overestimates the ones with higher scores.*

DATASET AUGMENTED WITH PAIRWISE MULTIPLICATIONS AND RATIOS

The original dataset has been augmented with pairwise multiplications and ratios. The covariates chosen for multiplications are: *citric.acid* and *total.slphur.dioxid* with a new covariate called *catsd; residua.sugar* and *pH* with a new covariate called *rspH; sulphates* and *alcohol* with a new covariate called *sa.* The covariates chosen for the ratios are: *fixed.acididty* and *chlorides* with a new covariate called *fac; density* and *pH* with a new covariate called *dpH; free.sulphure.dioxide* and *alcohol* with a new covariate called *fsda.*

The predictions with this model present the characteristic of overestimating the *quality* of wine with lower scores and underestimating those with higher scores.

The function *blasso* from the package *monomvn* has been once again applied to check if all the new variables created are really significant in predicting the *quality* of wine. The same values as before mentioned for the arguments *rd* and *ab* have been applied and the variable that can be dropped from the dataset are $\beta_3, \beta_5,\ \beta_7, \beta_9, \beta_{10}$ and β_{16} corresponding to the covariates *citric.acid, chlorides, total.sulphur.dioxide, pH, sulphates* and *fac.*

Predictions using this model have the same behaviour as almost all the models analysed previously: overestimates the *quality* of wine with lower scores and underestimates those with higher scores.

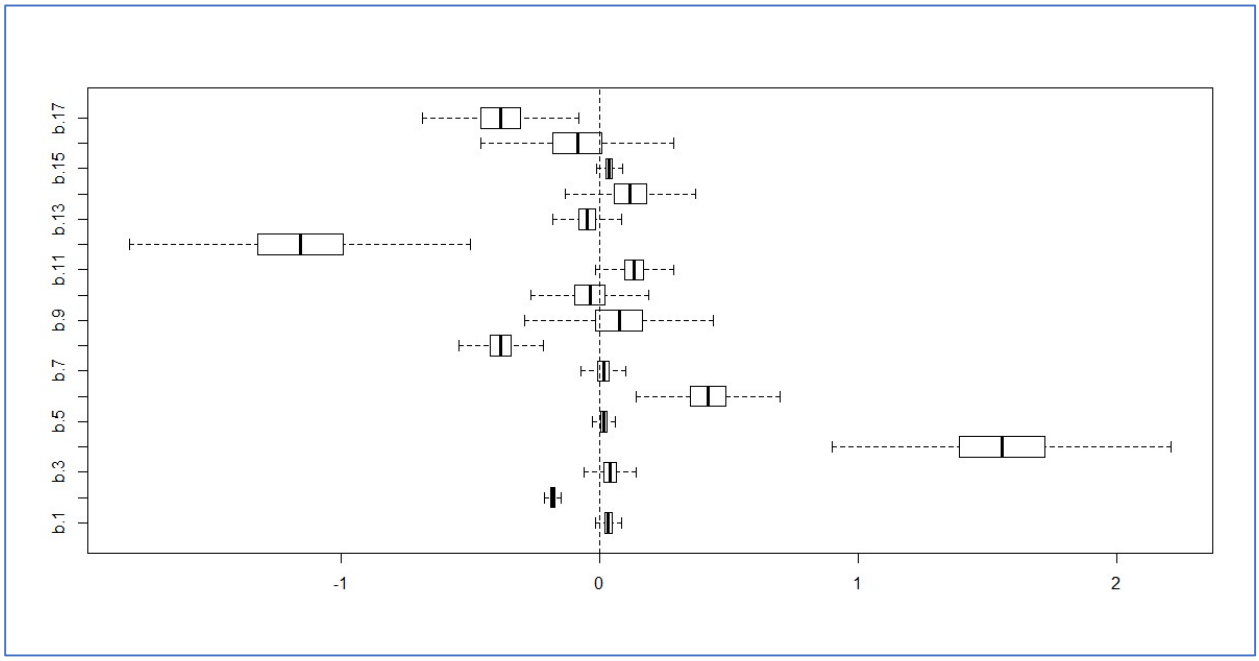

Figure 21 : *Central posterior intervals for regression coefficients β using Gibbs sampler.*

CONCLUSIONS

In conclusion, all the models created for the red wine dataset can be compared with each other. Same thing for the white wine dataset.

The function used for comparing different models is *loo* from the package *loo*. The function *loo* simulates a leave-one-out (loo) cross validation and one of the objects that returns is *estimates* which it contains a matrix with two columns and three rows. The columns are *estimates* and *se. estimates* are the actual *loo* estimates that are computed from the log-likelyhood matrix which has as row the total number of iterations in the posterior and as columns the total number of observations in the dataset. *se* is the standard error of the estimates. The three rows are *elpd_loo* which is the loo estimate*, p_loo* is the effective number of parameters in the model and *looic* is the loo estimate converted to a deviance scale and calculated as $-2 * elpd\_loo.$

Then the function *loo_compare* has been used which provides (*elpd_diff*) the difference between two or more models of the loo estimates along with the standard error (*se_diff*). If the difference is positive the second model is preferred, if the difference is negative the first model is preferred. The standard error gives information about the meaningfulness of this differences: if the absolute value of the *elpd_diff* is less than *se_diff* the two models compared do not perform differently, thus the simpler one should be chosen.

Table 6 and 7 reports all the *elpd_diff* values for all the models created for the red wine dataset and white wine dataset respectively. All the *elpd_diff* absolute values are bigger than the standard error, therefore, the loo estimates differences are meaningful, and the best models for both datasets are the ones created after applying a logarithmic transformation to the original dataset. Hence it seems that all the covariates are important in determine the *quality* of wine.

	ELPD_DIFF	SE_DIFF
Logarithmic Model	0.0	0.0
Model Selected	-2704.5	19.9
Augmented Selected Model	-2704.6	20.0
Augmented Model	-2707.7	20.2
Model_6	-2707.7	20.2

Table 5: *Comparison of all models created for the Red wine dataset. The preferred model is model created with the logarithmic transformation of the dataset.*
Logarithmic Model: model created with the logarithmic transformation of the dataset (page 11)
Model Selected: model created after applying the Bayesian lasso method to the original dataset (page 9)
Augmented Selected Model: model created after applying the Bayesian lasso method to the augmented dataset with pairwise multiplications and ratios (page 15)
Augmented Model: model created after augmenting the original dataset with pairwise multiplications and ratios (page 13)
Model_6: Model chosen for the best parameters of the priors (page 5)

	ELPD_DIFF	SE_DIFF
Logarithmic Model	0.0	0.0
Augmented Model	-8599.2	27.5
Augmented Selected Model	-8601.5	28.2
Model_3	-8622.0	28.2
	-8627.4	28.6

Table 6: *Comparison of all models created for the White wine dataset. The preferred model is model created with the logarithmic transformation of the dataset.*
Logarithmic Model: model created with the logarithmic transformation of the dataset (page 19)
Model Selected: model created after applying the Bayesian lasso method to the original dataset (page 17)
Augmented Selected Model: model created after applying the Bayesian lasso method to the augmented dataset with pairwise multiplications and ratios (page 19)
Augmented Model: model created after augmenting the original dataset with pairwise multiplications and ratios (page 19)
Model_3: Model chosen for the best parameters of the priors (page 17)

FUTURE IMPROVEMENT

This analysis can be improved by exploring in more details the most appropriate prior distribution to utilise. It could be also interesting to merge the two dataset and apply interactions with the two colours of the wine as a factor. Different distribution could also be taken into account for the response variable as the multinomial distribution and also exploring more transformation of the data such as the square root.