Bayesian Statistics Assignment

Judith Neve (0070661; j.a..nevedemevergnies@uu.nl)

Introduction

Taste is subjective and depends on individuals. However, it is generally agreed upon that certain products are of higher quality than others, such that some are considered luxury products while others are more affordable. Here, I focus on examining white wine quality, using information from an analysis examining red wine quality. I use two datasets: one containing information about 1599 red wines, and one containing information about 4898 white wines. Both datasets contain the same variables. The outcome variable of interest is wine quality, measured on a scale of 0-10. Each wine was assessed by at least three wine experts, and the median score of the assessment was taken to be the quality. The other variables are objective physicochemical measurements and are divided in two categories: direct impact on taste, which are easily described through adjectives relating to taste (e.g. “sour”, “sweet”), and no direct impact on taste.

Variables with a direct impact on taste are fixed acidity, volatile acidity, citric acidity, residual sugar, and alcohol content. Variables with no direct impact on taste are chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, and sulphates.

Research questions

I aim to determine which variables are good predictors of wine quality, assuming this is comparable across red and white wines. A preliminary analysis will be run on the red wine dataset to identify the least relevant predictors; those will be excluded. The remaining predictors will be examined in more detail using the white wine dataset. Three hypotheses are considered:

1. *Hypothesis 1*: Variables with a direct impact on taste are more important predictors of wine quality than variables with no direct impact on taste.
   This is because although people’s taste differ, wine experts are expected to have a more objective perspective on what wine is meant to taste like, and therefore variables with a direct impact on the taste will be more salient in the judgement of wine, while the way in which they impact the judgement will be fairly stable across experts.

2. *Hypothesis 2*: Variables with no direct impact on taste are more important predictors of wine quality than variables with a direct impact on taste.
   It is however also possible that experts, like anybody else, have their own preferences. In this case, variables that affect the wine in other ways than directly through taste may be more relevant for experts’ judgments.

3. *Hypothesis 3*: Both types of variables are similarly important.
   Of course, both taste and other types of measurements impact the experience one has when drinking wine. This hypothesis suggests a more holistic approach to evaluating wine quality.

   This will be examined using Bayesian multiple linear regression.

Methods

Model fitting

In order to compare the importance of predictors, all variables were standardised before any analyses were run. Using the red wine dataset, a frequentist linear regression (*Model 0*) was run using all variables as predictors of quality in order to gain prior knowledge to examine predictors of white wine quality. Predictors with a standardised coefficient with absolute value smaller than 0.1 were excluded from further analyses, as they were considered to have little predictive power of wine quality. The remaining predictors were included in the Bayesian regression models. Three models were fit:

1. *Model 1*: corresponding to *Hypothesis 1.* This model only included predictors with a direct impact on taste.
2. *Model 2*: corresponding to *Hypothesis 2.* This model only included predictors with no direct impact on taste.
3. *Model 3*: corresponding to *Hypothesis 3.* This model included all predictors considered to be relevant following fitting *Model 0.* This is the main model of interest.

The priors of the regression coefficients were Normal distributions, where the mean was set to be the regression coefficient of the corresponding parameter in *Model 0* and the standard deviation was set to be twice the corresponding standard error. These priors are detailed in *Table 1.* No intercept was included as the data was standardised and the intercept is therefore 0 by definition. Additionally, the error variance was set to have an inverse-gamma prior distribution with shape parameter $\alpha_0 = \frac{N_{red\ wine}}{2} = 799.5$ and scale parameter $\beta_0 = \sigma^2_{red\ wine} * \alpha_0 = 511.2$.

Table 1: Predictors included in Models 1-3 and priors of their regression parameters.

	Prior mean	Prior standard deviation	Model 1	Model 2	Model 3
volatile.acidity	-0.24	0.0537	x		x
chlorides	-0.11	0.0489		x	x
total.sulfur.dioxide	-0.13	0.0593		x	x
sulphates	0.19	0.048		x	x
alcohol	0.36	0.0699	x		x

The models were estimated using MCMC sampling. In *Model 1*, the regression parameter for volatile acidity was estimated using Gibbs sampling while the regression parameter for alcohol content was estimated using the Metropolis-Hastings algorithm. In *Model 2*, regression parameters for chlorides and total sulfur dioxide were estimated using Gibbs sampling and the regression parameter for sulphates was estimated using the Metropolis-Hastings algorithm. In *Model 3*, regression parameters for volatile acidity, chlorides, total sulfur dioxide, and sulphates were estimated using Gibbs sampling. The regression parameter for alcohol content was estimated using the Metropolis-Hastings algorithm. In all models, the error variance was estimated using Gibbs sampling.

The posterior distributions of the regression parameters drawn using Gibbs sampling were derived using conjugacy: they were determined to be Normal distributions with mean as shown in *Formula 1* and variance as shown in *Formula 2.* The posterior distribution of the error variance was determined to be an Inverse-Gamma distribution with shape parameter $\alpha = \frac{N}{2} + \alpha_0 = 3248.5$ and scale parameter $\beta = \frac{\sum_{i=1}^{N}(y_i - \sum_{j=1}^{5} b_j^{(t)} * x_{ij})^2}{2} + \beta_0$ at the t^{th} iteration of the sampler.

Formula 1. mean of the posterior distribution of the regression parameter j.
$\frac{\frac{\sum_{i=1}^{N} x_{ij} * (y_i - \sum_{k \neq j}(b_k^{(t)} * x_{ik}))}{\sigma^{2(t-1)}} + \frac{prior\ mean}{prior\ variance}}{\frac{\sum_{i=1}^{N} x_{ij}^2}{\sigma^{2(t-1)}} + \frac{1}{prior\ variance}}$, where x_{ij} is the i^{th} observation of predictor j, y_i is the i^{th} observation of wine quality, *b* are the most recent draws of the regression coefficient (so for volatile acidity, all *b* are from the previous iteration of the chain, while for chlorides, the *b* for volatile acidity is from the current iteration

but the others are from the previous iteration, etc.), and $\sigma^{2(t-1)}$ is the error variance from the previous iteration.

Formula 2: variance of the posterior distribution of the regression parameter j

$$\frac{1}{\frac{\sum_{i=1}^{N} x_{ij}^2}{\sigma^{2(t-1)}} + \frac{1}{prior\ variance}}$$

The Metropolis-Hastings algorithm was used to sample values for last regression coefficient of each model. This step took place after all other regression coefficients had been sampled within the iteration, but before sampling the residual variance. The proposal density that was used was the same as the prior density.

For each model, two chains of 10,000 iterations with a burn-in period of 500 were run. Starting values were drawn at random from a Uniform(-1, 1) distribution for the regression coefficients and from a Uniform(0, 1) distribution for the error variance. The bounds of the uniform distributions were chosen due to the data being standardised, thus leading the parameters to be theoretically bound within those intervals. Model convergence was assessed using trace plots, autocorrelation plots, the Gelman-Rubin statistic, and the MC error.

Regression assumptions

Linear regression assumes all predictors have a linear relationship with the outcome. This assumption was tested using a posterior predictive check. The test statistic was defined to be the number of predictors with an absolute correlation of > 0.2 with the outcome. This test statistic is powerful in that, would a predictor be very weak, simulated data may display a weaker correlation with the predictor. Counting the number of correlations allows for variation in the strengths of the linear relationships of predictors with the outcome as each predictor is considered separately. This assumption was only checked for *Model 3*.

Model selection

Two measures were used to select the best model and hypothesis. One was the DIC: it was computed for *Model 1*, *Model 2* and *Model 3*. The DICs were compared to determine the best model, i.e., the model with the smallest DIC. The other was the Bayes Factor: using the estimates from *Model 3*, Hypothesis 3 (unconstrained) was compared with both Hypothesis 2 and Hypothesis 1 to determine which had the most support.

Results

Models 1 and 2

Convergence was reached for both of these models. Convergence checks and parameter estimates and interpretation are not detailed here as these models were only fit in order to compare them with *Model 3* using the DIC.

Model 3

Convergence appears to have been reached. Trace plots for all parameters are presented in *Figure 1*. Autocorrelation plots are shown in *Figure 2*.

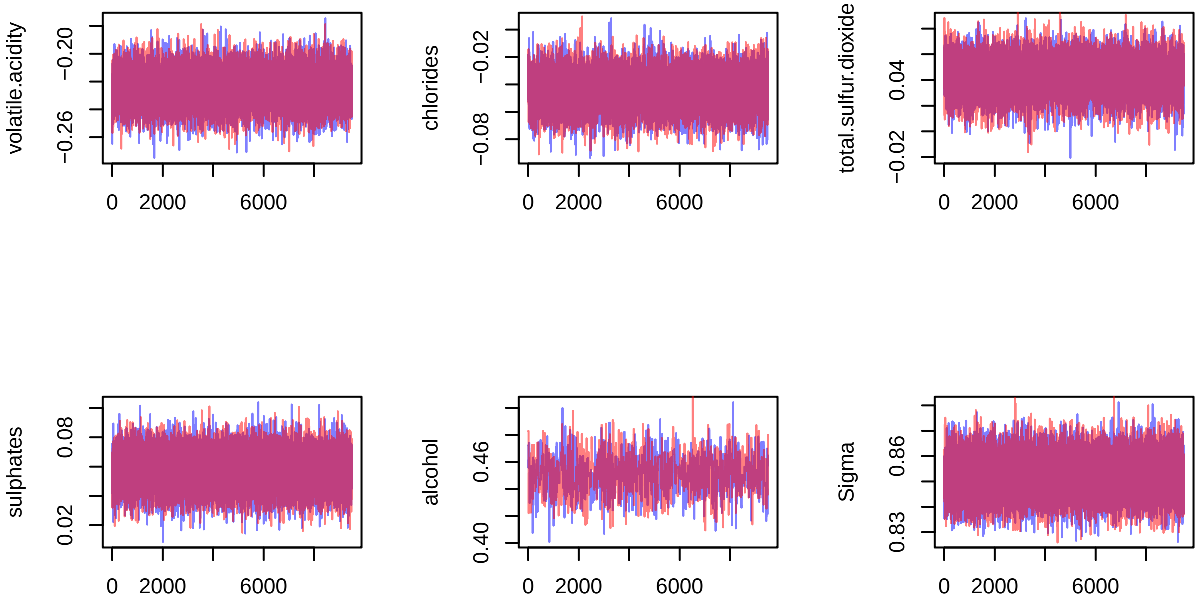

Figure 1. Trace plots for model parameters in Model 2.

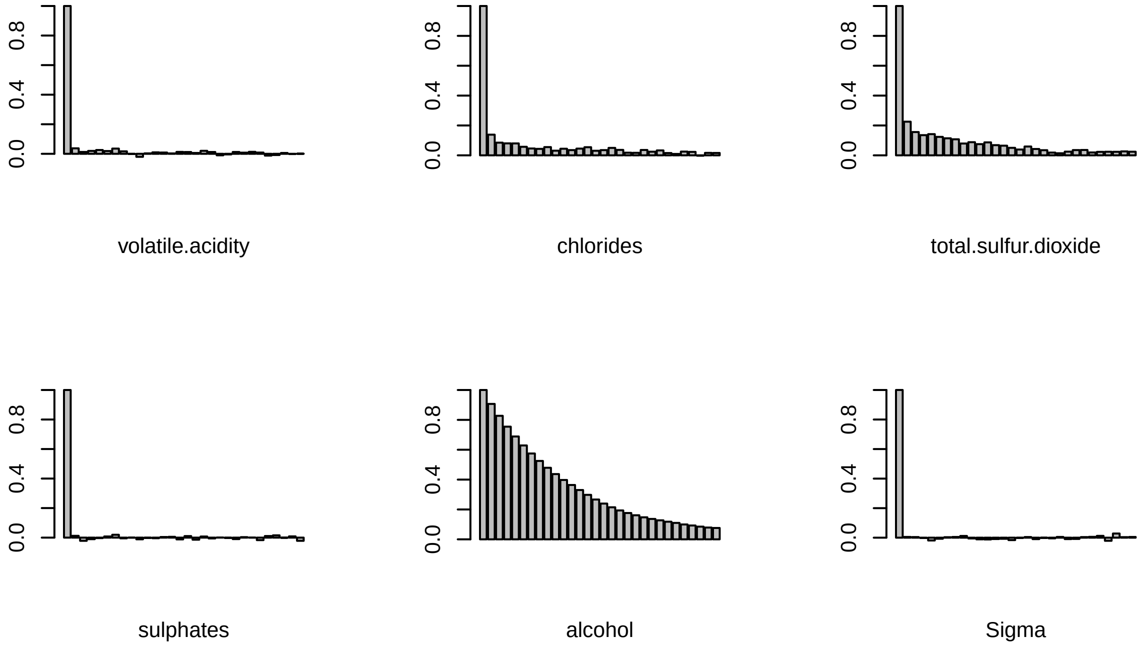

Figure 2. Autocorrelation plots for the first chain of Model 2.

Alcohol content shows less variation and much higher autocorrelation than the other two as it was sampled using Metropolis-Hastings and the acceptance rate was low (0.101 overall). However, we see from the trace plot that it still varied a fair amount and appears to have converged. Further, we do not worry about this as the Gelman-Rubin statistics of all parameters were extremely close to 1 (largest absolute deviation from 1: 0.00006) and the MC errors were all smaller than 0.00015, while 5% of the smallest standard deviation of the sampled parameters was 0.0005, thus the MC errors are all well below their suggested maximum values.

The ppp-value for the assumption of linear relationships was 0.4959, which is extremely close to 0.5 and thus shows strong support in favour of the assumption not being violated.

Parameter estimates of *Model 3* are presented in *Table 2*. It appears alcohol content is the strongest predictor of quality, where for a one standard deviation increase in alchohol content, quality can be expected to increase by 0.4517 standard deviations on average. The 95% credible interval for it is (0.4243; 0.4783), that is, there is a 95% probability that the true value of the regression coefficient lies within these bounds. As this interval is far from overlapping with any other interval, it is clear alcohol content is the strongest predictor. The weakest predictors is total sulfur dioxide, where for a one standard deviation increase in total sulfur dioxide, quality is expected to increase by 0.042 standard deviations on average. The residual standard error is estimated to be 0.85, indicating a lot of variance in wine quality remains unexplained.

Table 2: Model 3 results.

	Posterior mean	95% Credible Interval
volatile.acidity	-0.2247	(-0.2482; -0.201)
chlorides	-0.0448	(-0.0696; -0.0202)
total.sulfur.dioxide	0.042	(0.0152; 0.0684)
sulphates	0.0574	(0.0339; 0.0806)
alcohol	0.4517	(0.4243; 0.4783)
Sigma	0.8521	(0.8378; 0.8669)

The DIC of *Model 3* was 12528, while the DIC of *Model 1* was 12563 and the DIC of *Model 2* was 13577. This indicates that the model with the mix of predictors is the best model.

The Bayes Factor of *Hypothesis 1* against *Hypothesis 3* was 15.14, such that it appears there is over 15 times more support for the hypothesis that taste predictors are stronger predictors of wine quality than there is for the hypothesis that quality is not predicted more by one category of predictors than another. The Bayes Factor of *Hypothesis 2* against *Hypothesis 3* was 0, showing no support for the hypothesis that predictors not directly related to taste are stronger predictors of wine quality than predictors directly related to taste.

Taking these results in conjunction with the DIC, we may conclude that while taste-related predictors are the strongest predictors of quality, they are not the only important predictors and the other predictors should also be included in the model.

Discussion

I investigated which predictors are most relevant to predict wine quality, comparing three hypotheses: whether these predictors are primarily taste-based, primarily not taste-based, or whether both types of predictors are relevant to predict wine quality. Strong support in favour of *Hypothesis 1* (variables with a direct impact on taste are stronger predictors of wine quality) was shown using the Bayes Factor to compare all three hypothesis using estimates from *Model 3* (the model with all predictors deemed relevant following the preliminary analysis). However, when comparing the fuller model with models only including one category of predictors, it appeared that entirely removing the predictors that are not taste-based led to a worse model. I therefore concluded that both types of predictors are necessary to predict wine quality, but the taste-based ones are more important. I also found that predictive power is limited, as the estimated error variance remained fairly large.

This analysis was conducted using Frequentist statistics to acquire prior knowledge and Bayesian statistics to evaluate hypotheses. The hypotheses were examined with more nuance using Bayesian statistics than they would have been using Frequentist statistics: Frequentist approaches would only have allowed for the comparison of hypotheses using models including or excluding certain predictors, or would have required the creation of composite measures for each type of predictor. Using a Bayesian approach allowed for both testing the necessity to include both categories of predictors, and for the difference in importance between the types of predictors.

Additionally, the Bayesian approach allowed me to test an assumption of linear regression more formally than Frequentist statistics would have: in a Frequentist approach, this assumption would have been assessed via a visual examination of a scatterplot for each parameter. This would have been more subjective, and would only have provided individual conclusions regarding each predictor. Using a Bayesian approach allowed me to combine the information from all predictors in order to assess the assumption for the overall model.