Final Report for Red Wine Analysis

Ruiqiang Chen, Michael DeWitt, David Williams, Alex Vannoy

7/28/2017

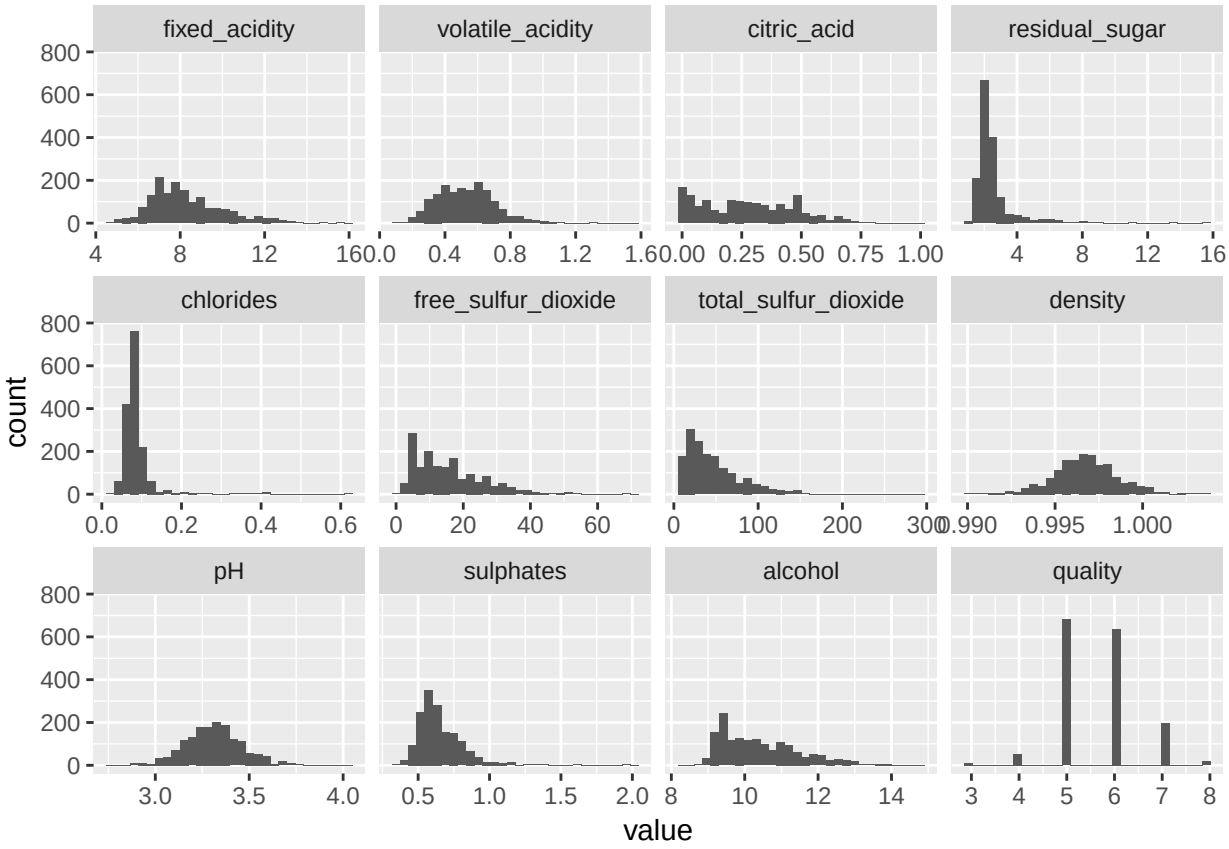
Introduction

The purpose of this document is to report the proposed statistical models for classification of red wine.

Description of Data

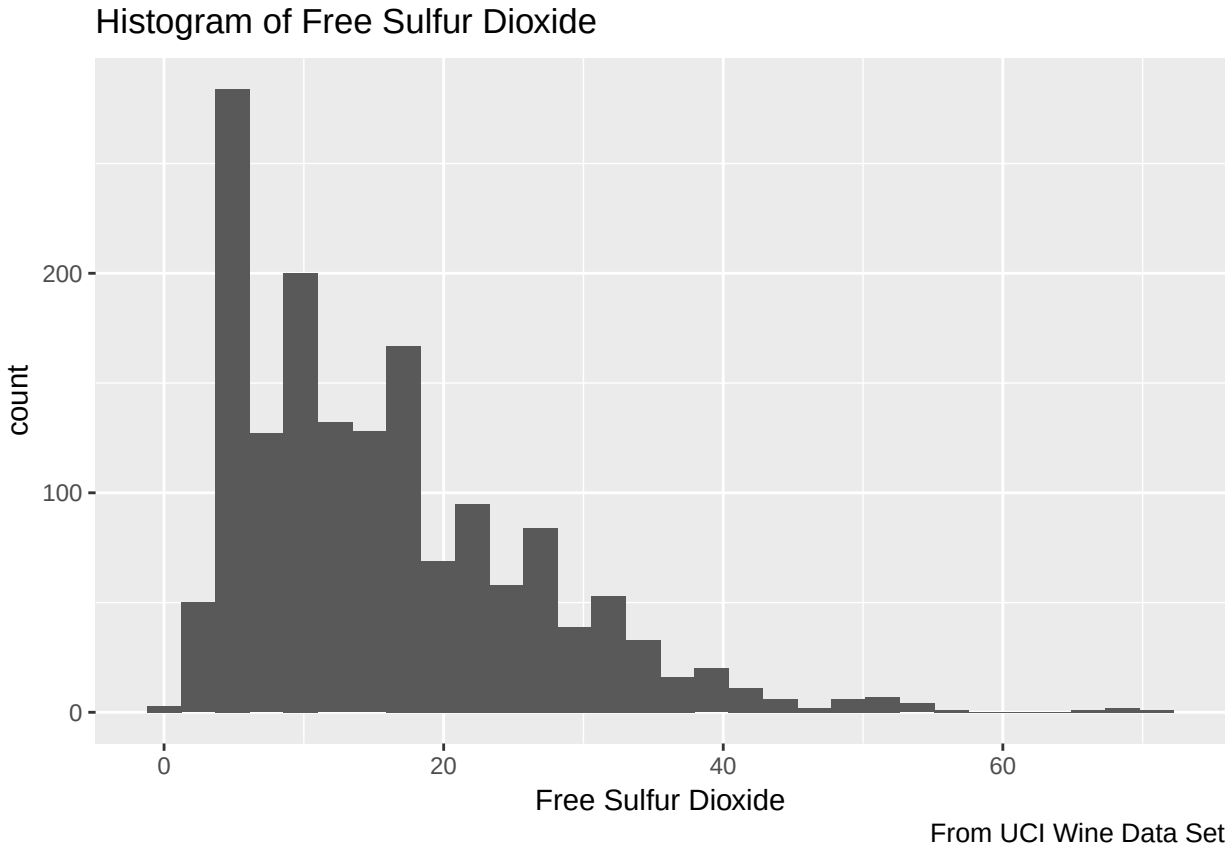
The data set provided is the Wine dataset from UC Irvine. It consists of 1599 with a total of 12 predictors. These predictors include the following fixed_acidity, volatile_acidity, citric_acid, residual_sugar, chlorides, free_sulfur_dioxide, total_sulfur_dioxide, density, pH, sulphates, alcohol, quality with the quality feature being associated with the judgement of the individual wine's quality. Quality is the feature of interest for the dataset as the vinter is interested in judging the wine's quality through objective means rather than todays subjective method of averaging the 1-10 point judgment of tastetesters. The distribution of these different criteria can be seen below:

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Reviewing the individual components there appears to be a slight irregularity with total free sulfur dioxide. This can be seen in the histogram of this variable.

```
## Using classification as id variables
```



Method

Regression

In order to select the best fit regression model several different modeling methods were tested. These include LEast Squares Regression, Ridge Regression, Lasso Regression, Principle Components Regression and Partial Least Squares Regression. For each of these methods the quality integer was the value that the model was attempting to predict. The data was divided into two sets, a training set to train the model and a testing set for model validation. We will now go deeper in the model generation process for each of these different modeling types and methods.

Least Squares

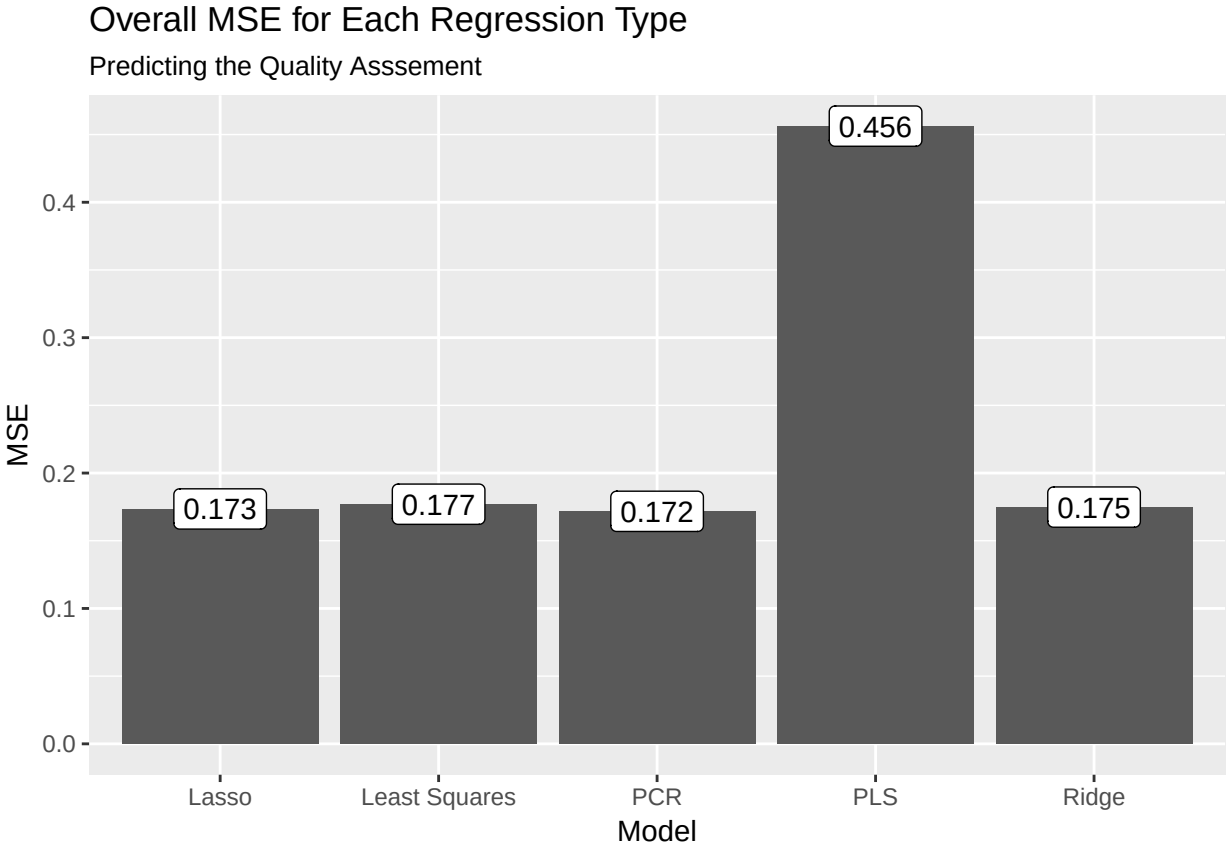
Ridge Regression

Lasso Regression

Principle Components Regression

Partial Least Squares Regression

Model Selection



Residual Analysis

Classification

Model Selection

Residual Analysis

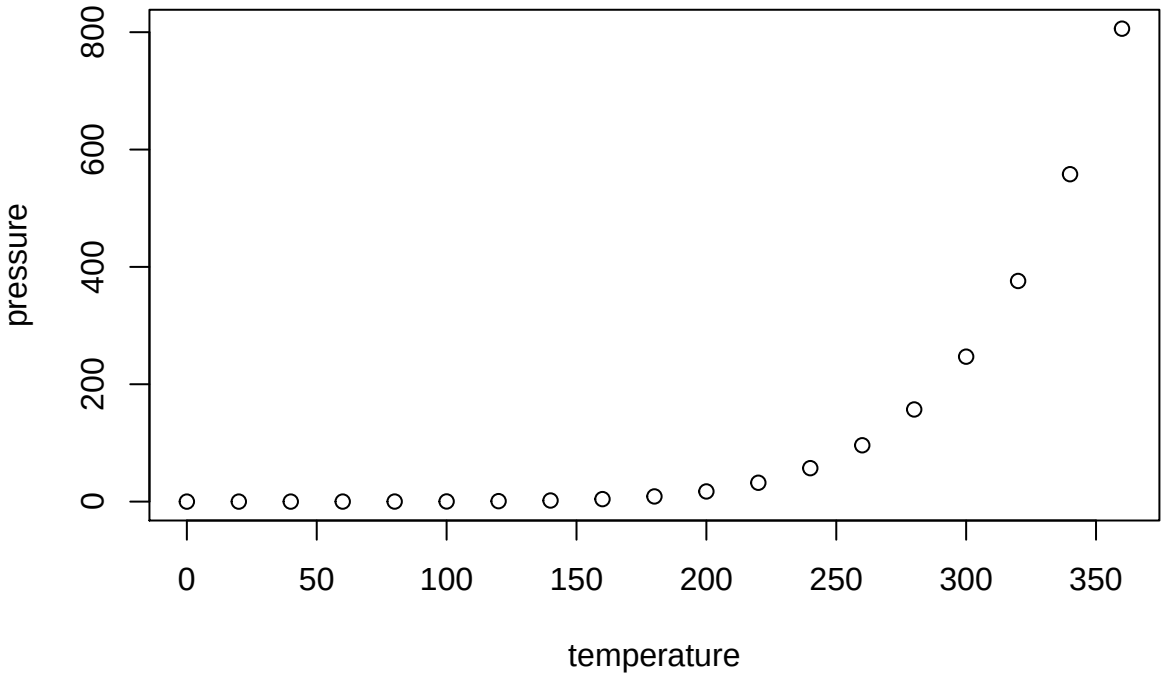
Comparison of Models

Discussion

Conclusion

Issues

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.