Wine Quality Analysis

Stat 3340

Group 33

Author:

Chu Yifei B00765157

Yasong Wang B00756812

Ruming Liu B00811402

Dalhousie University

Abstract

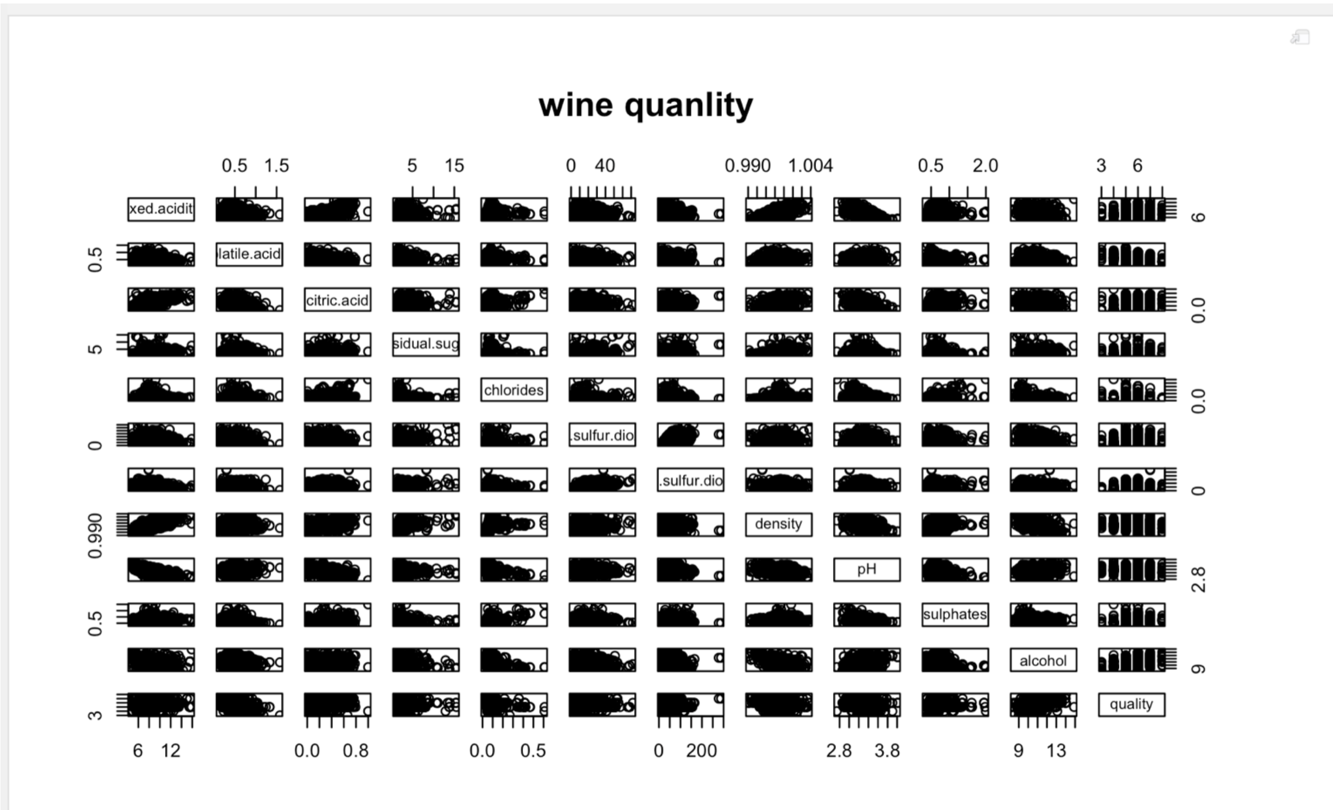
This project emphasizes the regression analysis performed on wine quality data. The variable of interest here is the wine quality. First, a linear regression model is fit using wine quality as the dependent variable and all the other variables as an independent. This model did not provide a good R squared. To verify this, only the significant variables were left in the model, and the rest of the variables were removed. This did lead to a drastic increase in the R squared value and, hence providing a better model.

Introduction

The wine quality dataset contains 12 variables with 1599 observations. The question of interest is, we want to find the line of best fit for the quality of wine using the other explanatory variables. We wish to estimate the quality of red wine using the linear regression model.

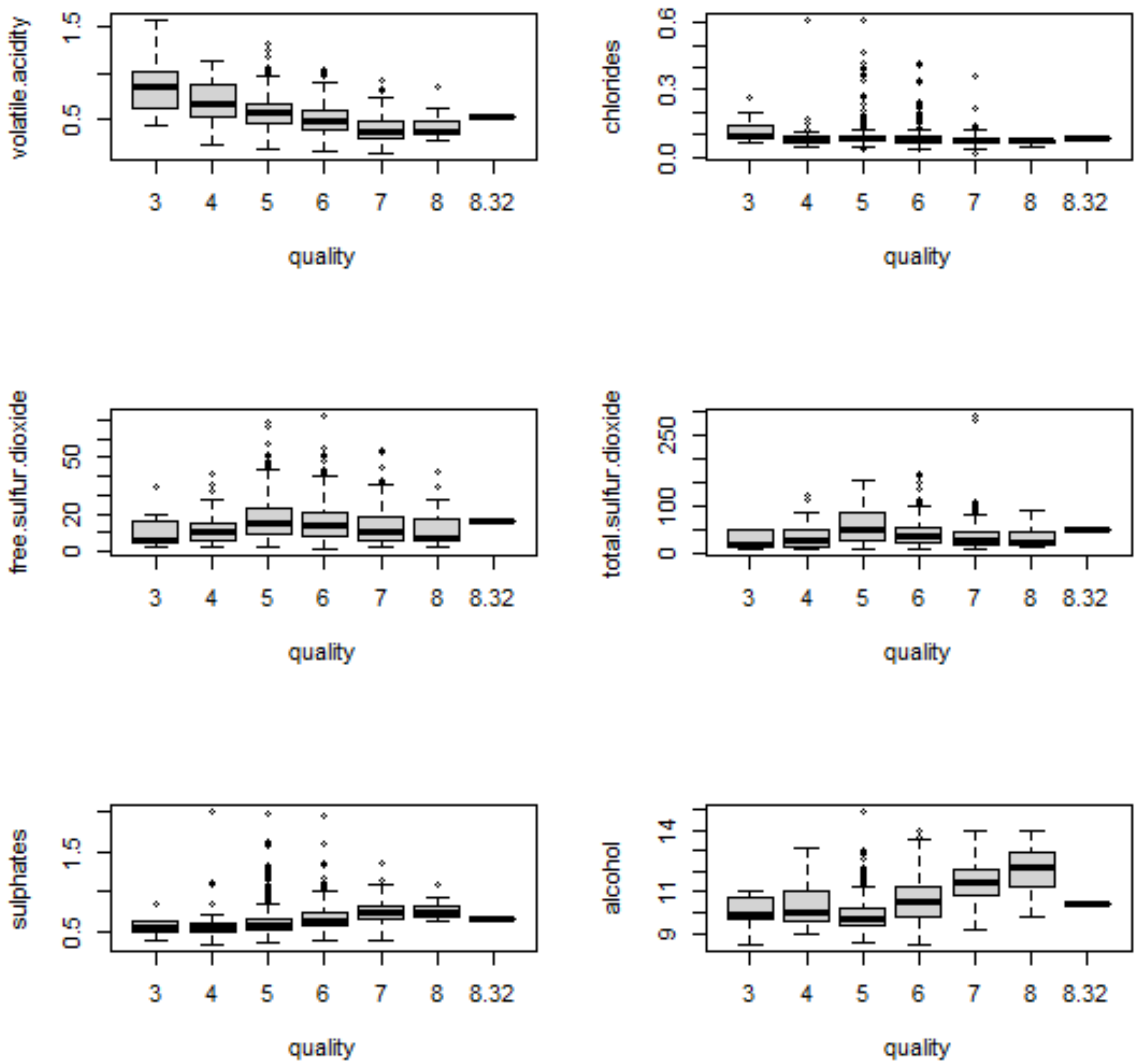
Data Description

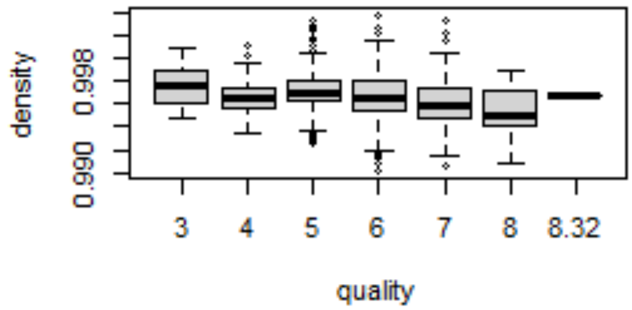
The dataset contains 11 numeric explanatory variables and wine quality having six categories starting from 3 to 8, the five-point summary of explanatory variables are given below,

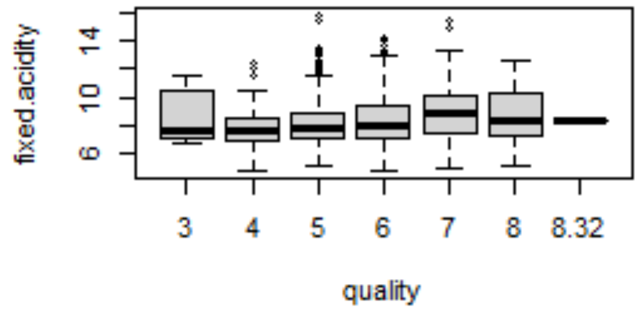


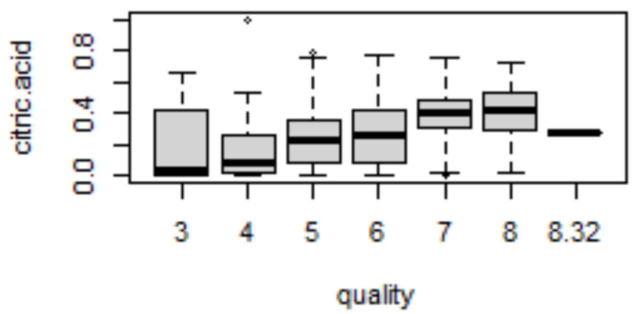

```
 fixed.acidity   volatile.acidity  citric.acid    residual.sugar  
 Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900  
 1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900  
 Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200  
 Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539  
 3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600  
 Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500  
   chlorides       free.sulfur.dioxide total.sulfur.dioxide    density      
 Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901  
 1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956  
 Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968  
 Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967  
 3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978  
 Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037  
       pH          sulphates         alcohol     
 Min.   :2.740   Min.   :0.3300   Min.   : 8.40  
 1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50  
 Median :3.310   Median :0.6200   Median :10.20  
 Mean   :3.311   Mean   :0.6581   Mean   :10.42  
 3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10  
 Max.   :4.010   Max.   :2.0000   Max.   :14.90  
```

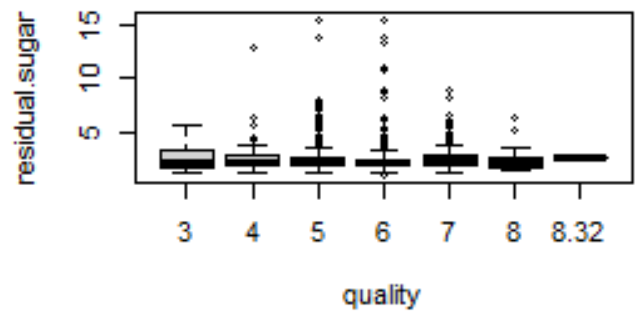
And the Box-plot for all the explanatory variables by wine quality is given below. The points outside the hinges of the box-plot are outliers. From the box plots, it is clear that the data is approximately normally distributed.

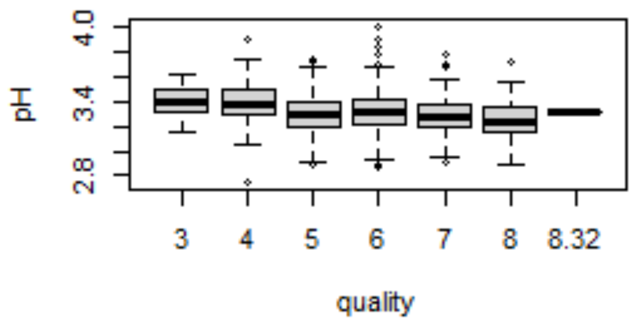












The new data point is added by using the command,

dataset <- rbind(dataset,c(8.32,0.5278,0.271,2.539,0.0874,15.87,46.47,0.9967,3.311,0.6581,10.42,6))

The new data points are obtained from the summary statistic mean, i.e. the data point for all the variables is mean value and median is used for quality.

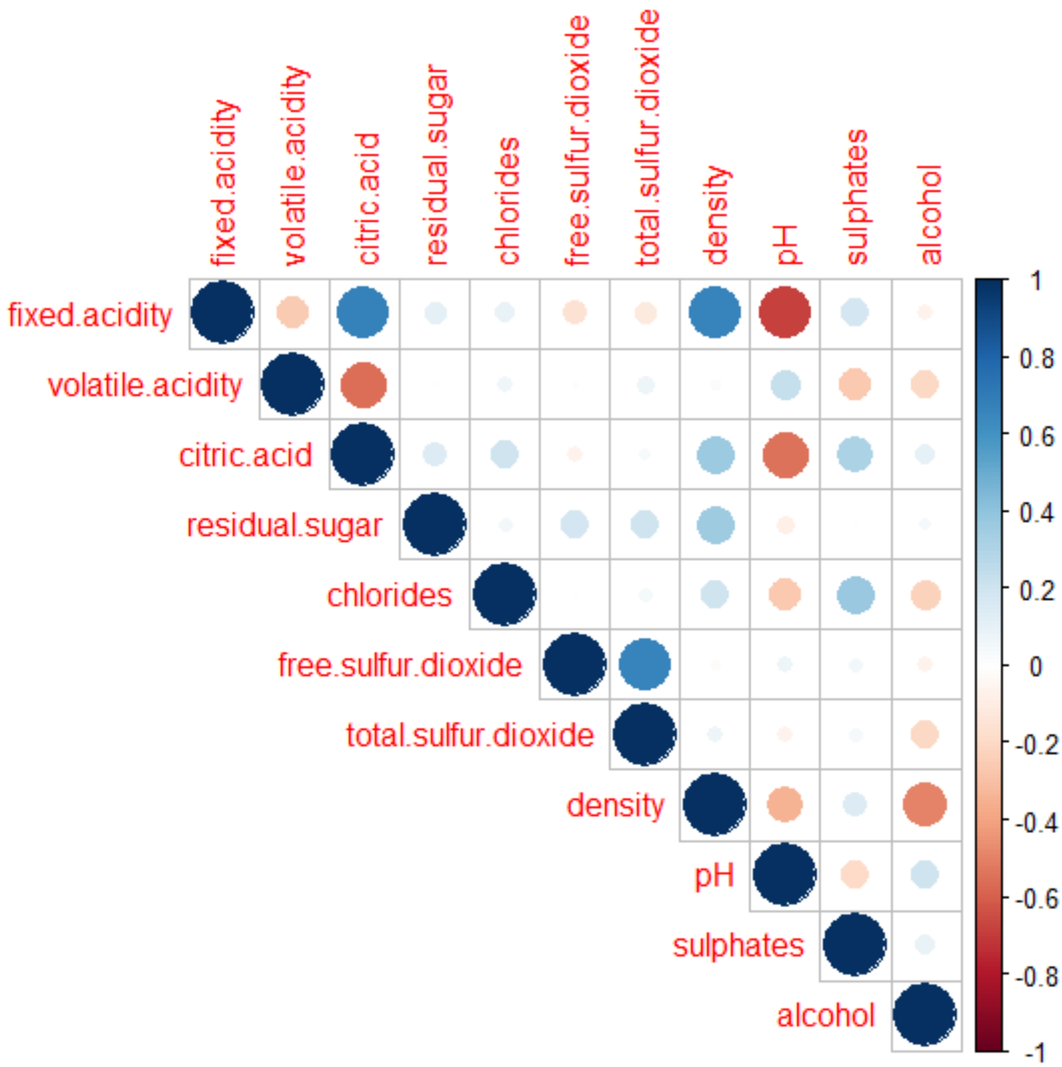
Correlation Matrix and Plot

Before fitting the linear regression model, we compute a correlation matrix for the explanatory variables and plot the results. The correlation matrix is given by, And the Box-plot for all the explanatory variables by wine quality is given below. The points outside the hinges of the box-plot are outliers. From the box plots, it is clear that the data is approximately normally distributed. And there exist several outliers in the dataset.

	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide
fixed.acidity	1.00	-0.26	0.67	0.11	0.09	-0.15
volatile.acidity	-0.26	1.00	-0.55	0.00	0.06	-0.01
citric.acid	0.67	-0.55	1.00	0.14	0.20	-0.06
residual.sugar	0.11	0.00	0.14	1.00	0.06	0.19
chlorides	0.09	0.06	0.20	0.06	1.00	0.01
free.sulfur.dioxide	-0.15	-0.01	-0.06	0.19	0.01	1.00
total.sulfur.dioxide	-0.11	0.08	0.04	0.20	0.05	0.67
density	0.67	0.02	0.36	0.36	0.20	-0.02
pH	-0.68	0.23	-0.54	-0.09	-0.27	0.07
sulphates	0.18	-0.26	0.31	0.01	0.37	0.05
alcohol	-0.06	-0.20	0.11	0.04	-0.22	-0.07

	total.sulfur.dioxide	density	pH	sulphates	alcohol
fixed.acidity	-0.11	0.67	-0.68	0.18	-0.06
volatile.acidity	0.08	0.02	0.23	-0.26	-0.20
citric.acid	0.04	0.36	-0.54	0.31	0.11
residual.sugar	0.20	0.36	-0.09	0.01	0.04
chlorides	0.05	0.20	-0.27	0.37	-0.22
free.sulfur.dioxide	0.67	-0.02	0.07	0.05	-0.07
total.sulfur.dioxide	1.00	0.07	-0.07	0.04	-0.21
density	0.07	1.00	-0.34	0.15	-0.50
pH	-0.07	-0.34	1.00	-0.20	0.21
sulphates	0.04	0.15	-0.20	1.00	0.09
alcohol	-0.21	-0.50	0.21	0.09	1.00

The above matrix can easily be understood using the correlation plot given below,



Here, we can see that the darker regions of blue and red color represent significant correlation. It is advised to use only those explanatory variables which are uncorrelated with each other. It is clear that fixed.acidity is correlated with many other explanatory variables, including it will result in the problem of multicollinearity.

Methods

Multiple Linear Regression Model: The multiple linear regression model is of the form,

$$y_j = b_0 + b_1 x_{1j} + b_2 x_{2j} + \cdots + b_k x_{kj} + \in$$

Where y is the dependent variable present in the study, xj's are the independent or explanatory variables present. bi's being the coefficients of the independent variables. Is the error term, normally distributed with mean 0 and a constant variance.

The linear regression model can be applied using the *lm()* function in R, and model summary can easily be obtained by *summary(model_name)* function. If in the regression model, the p-value for the significance of the model is less than 0.05, then the model is assumed to be significant at 95% confidence level. Also, the p-value should be greater than 0.05 for the individual significance of the explanatory variables

The Regression Diagnostics can be done using the plot() function, which gives the plots for linearity, common variance, normality and outlier values. The assumption of multicollinearity can be checked using the correlation matrix of the explanatory variables used in the model. The correlation matrix can be obtained using the command rcorr*()* under the library **Hmisc,** and the plot can be obtained using *corrplot()* command under the library **corrplot.**

Results:

The linear regression model was fit on the wine quality data. The dependent variable was the quality and

The rest of the variables were independent variables. First, a regression model was applied, taking into account all the explanatory variables. Then removing the variables that were not significant, another regression model was built using the lm() function.

The R square obtained was 0.36 when all the significant explanatory variables were considered in the model. This implies that only approximately 35% variation in the model was explained by the variables that were used in the model.

A R square value of 0.7 or greater is considered to be an extremely good model and any value of R square less than 0.3 is not considered as a very good model.

Conclusion:

The conclusion drawn from the study is that using a linear regression model does not provide extremely accurate results when predicting the wine quality. Also, it is to be noted that as the wine quality variable is measured on an integer scale, it would have been better to use some classification technique such as a multinomial logistic regression model in order to accurately classify the variables.

.