Multiple Linear Regression model on Wine Dataset

3/29/2020

Introduction

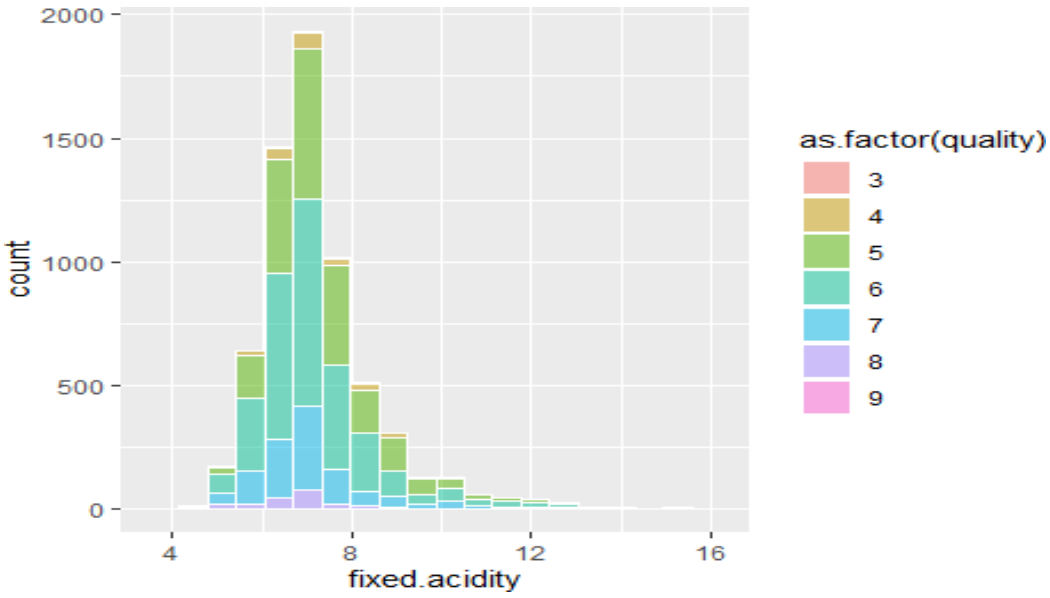
First of all, we downloaded red and white wine datasets and merged them together as a single dataset. In this report we are going to predict the quality of wine by considering one or more variables in our regression model.This merged dataset "Wine" contains 6497 observations and 12 variables which gives information mainly about the composition of red and white wines.

Data and Planning

Combining datasets for our analysis

By looking at our new dataset "Wine" (From summary in rmd file) we can say that there are no missing values in our data. Here our outcome variable is quality and remaining 11 variables are our input variables. Lets visualize the data to identify the distribution and frequency of a particular variable in wine dataset (eg: we considered fixed acidity variable as a factor of quality variable) by plotting a histogram.

Exploratory data analysis



As from the above graph, we can say that the dataset has the observations where the quality of wine varies from 3 to 9. We can clearly notice that most of the observations are for '5' and '6' quality wines (we can confirm this from the code in rmd file).

Selecting the predictor variables

We plotted the boxplots (from rmd file) and we found that, '5' and '6' quality wines are having high values but we cannot say that these are outliers. The means of volatile acidity, total sulphur dioxide, density, chlorides, pH and alcohol looks significantly different when we consider '5' and '6' quality wines. Therefore we can consider these variables as significant variables for determining our outcome variable quality. Let's confirm our predictor variables by perforiming linear model on all our variables to find the significant variables.

The results of linear model on quality (from rmd file) confirm that fixed acidity, volatile acidity, residual sugar, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol are significant variables.

Checking correlation for input variables

From the correlation plot (in rmd file), we can say that free sulfur dioxide and total sulfur dioxide are correlated with each other and alcohol is correlated with density.

1. We can also say that there is no high correlation between the predictor variables i.e. so we continued with the assumption of no perfect multicollinearity.

2. There are no external variables. These are the only variables under consideration.

Multiple linear regression by considering the significant variables

We have selected the predictor variables for our model by using all subsets methods.Here we have 11 input variables, after finding the significant variables we used subset method and found 4 input variables alcohol, sulphates,volatile acidity,free sulfur dioxide (by trail and error) as best predictors for our model. (Refer model in rmd file)

Checking Assumptions for our model

Variable types:

All the predictor variables are quantitative and the outcome variable quality is also quantitative.

Non-zero variance:

The predictors are having some variance, which can be observed from the graphs.

No perfect multicollinearity

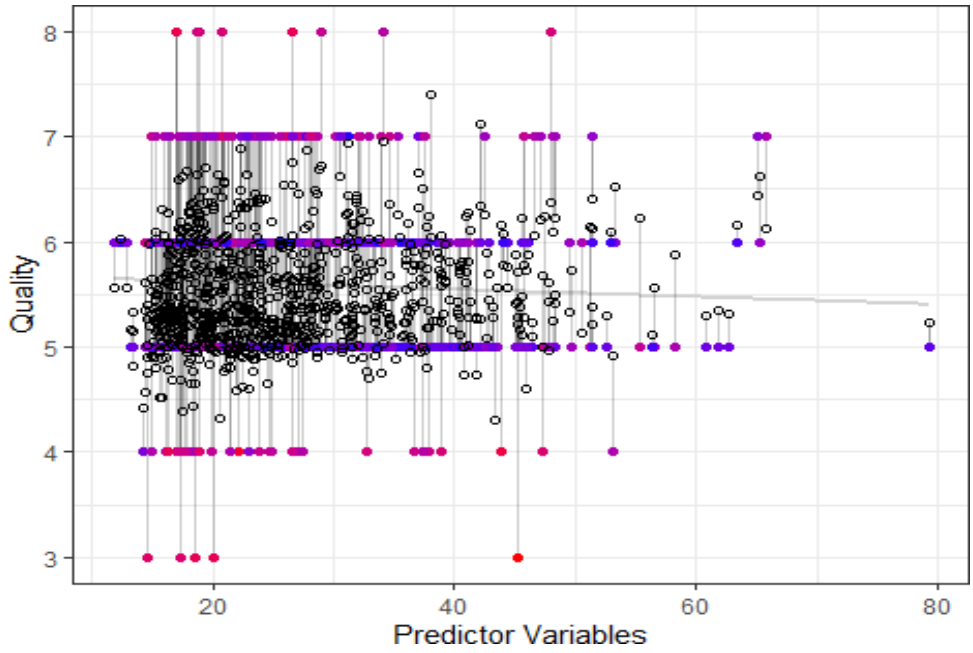
The predictors are not higly correlated with quality variable i.e. 'R' value (correlation coefficient) is not more than 0.80 or 0.90 [From text book 7.7.2.4. Multicollinearity].We checked VIF to confirm the multicollinearity assumption.The largest VIF was 1.210464, less than 10. The average VIF was 1.12921 which is close to 1. Therefore, we conclude that there is no collinearity in our data.

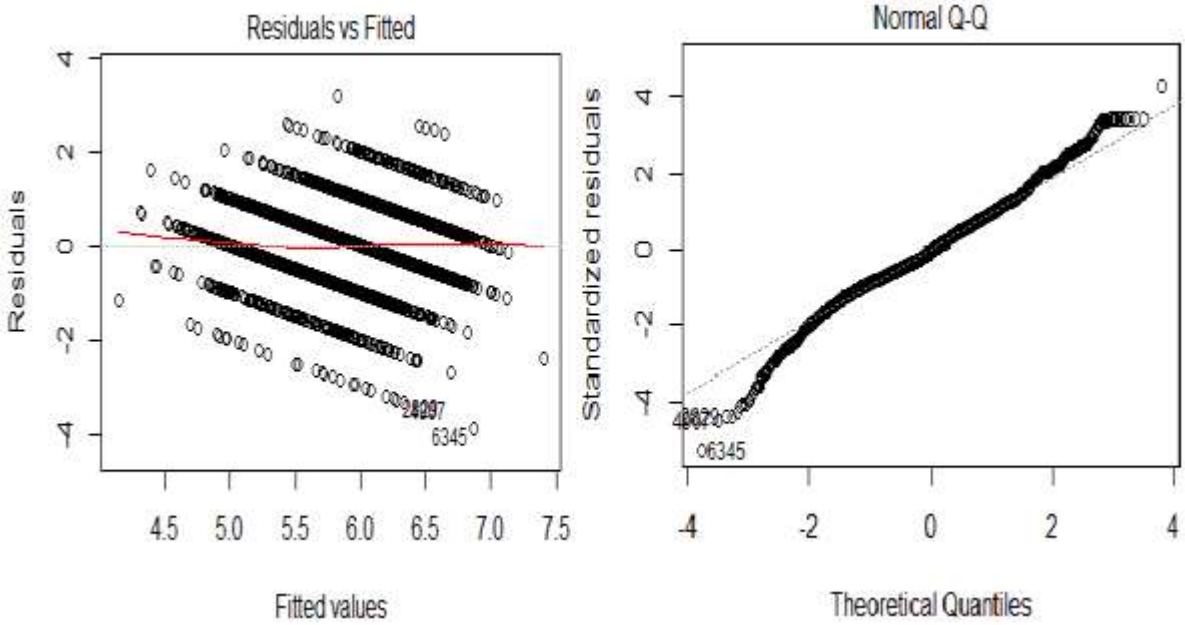
Independent Errors

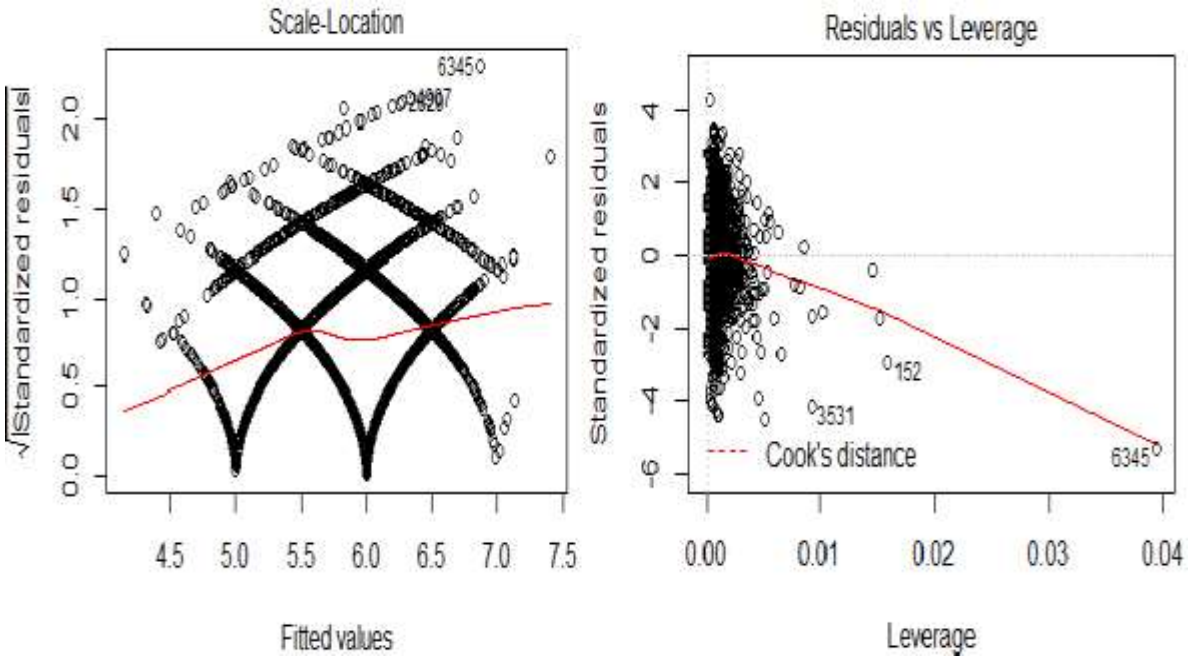
The Hypotheses for the Durbin Watson test are: H0 = no first order autocorrelation. H1 = first order correlation exists. The Durbin-Watson test for independent errors was significant at the 5% level of significance (d=1.65, p=0).So, we reject the null hypothesis.Therefore, we can say that the residuals

are correlated and we are not meeting this assumption.As a rule of thumb the test statistic value is 1.65 which is in the range of 1.5 to 2.5 and this states that they are relatively normal.

Testing residuals for homoscedastic and normally distributed errors







From the above plots, we can say that the residuals are not clearly linear but they are homoscedastic i.e. the variance is constant. Q-Q plot also looks normal.

Influential Cases
we calculated Cook's Distance on our model. Cook's distance was a maximum of 0.2304569, far below the chosen cutoff value of 1. So, we can conclude that there are no influential cases.

Checking for possible outliers

We found 392 residuals are above or below 1.96 standard deviations. As this represents 6.03% of the observations, expected if the residuals are normal (5% of data is expected to be outside of 2 standard deviations), we do not consider any of these observations as outliers and continued with all 6497 observations included in the model.

Evaluation of the Model

```
##   quality predicted    residuals
## 1       5  4.936541  0.063459134
## 2       5  5.059420 -0.059419777
## 3       6  5.656103  0.343897267
## 4       5  4.996190  0.003810348
```

Conclusion

All the predictor variables have an influence on quality of wine at the 5% level of significance (from summary of the model).This model concludes that Alcohol, Sulphates, Volatile acidity and free sulfur dioxide explains 27.31% of the variance in predicting the wine quality.We have evaluated our model by predicting some observations which can be seen in the table above.From the results, we can say that the model is predicting accurately but not 100%. This is the limitation of this model. This is because; there might be some external variables which are not included in the dataset that is affecting the prediction of quality.