Linear Statistical Models

SYS 6021

Project 4

Wine Quality Prediction

Debajyoti Datta
dd3ar@virginia.edu

Honor Pledge: On my honor, I pledge that I am the sole author of this paper and I have accurately cited all help and references used in its completion.

Summary

Wine quality varies greatly according to a variety of things. Starting with natural factors like average growing season temperature during harvesting to the amount of rainful. It can also be very intricate such as the amount of winter rainful or summer rainful. It can also vary greatly in terms of the chemical composition of the wine such as the amount of sulphur dioxide, acidity, chlorides and so on. These various aspects have been analyzed for the prediction of the quality of wine based on generalized linear models and through bootstrapping and regression for the analysis of the price. The final model for the prediction of the wine prices had a p-value of 1.359e-06 and the adjusted $R\hat{2}$ of 0.7185. The test set price prediction is quite good and the root mean squared error on the test set is 0.08199167 which shows that the price depends significantly on these factors. The final model for the prediction of wine quality has an accuracy of 0.84 and a precision of 0.84, suggesting that a lot can be predicted about wine quality and wine prices based on various input features like chemical composition and other natural factors before brewery.

1. Problem Description

1.1 Situation:

Once viewed as a luxury good, nowadays wine is increasingly enjoyed by a wider range of consumers and is gaining international attention. Vinho Verde is a Portuguese wine that originated in the north of the country. As a result of such a upwards trend in quality, Vinho Verde is seeing a growth in the value of its exports, with sales to the UK, its third largest export market after the US and Germany, up 20% from 1.3 million in 2012 to 1.5m in 2013. In the first half of 2014, sales abroad (outside Portugal) were 27.3 million, against 23.9 million a year earlier. To support its growth, the wine industry is investing in new technologies for both wine making and selling processes.

With so much popularity of this variant, it seems like an interesting task to be able to predict its quality, or rather human wine taste preferences using a set of independent variables affecting it, including, fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol. For this task quite a few models are being developed and compared so as to see how the performance of the model can be best measured and if the quality of wine can really be predicted from the features mentioned here.

The data set chosen has a collection of 4898 observation of 12 variables contributing to the quality of the white variant of Vinho Verde. The wine has been graded by experts on a scale of 0 to 10, 10 being the highest quality and 0 being the worst. The output is therefore based on sensory data, while the input is a result of objective, physicochemical tests.

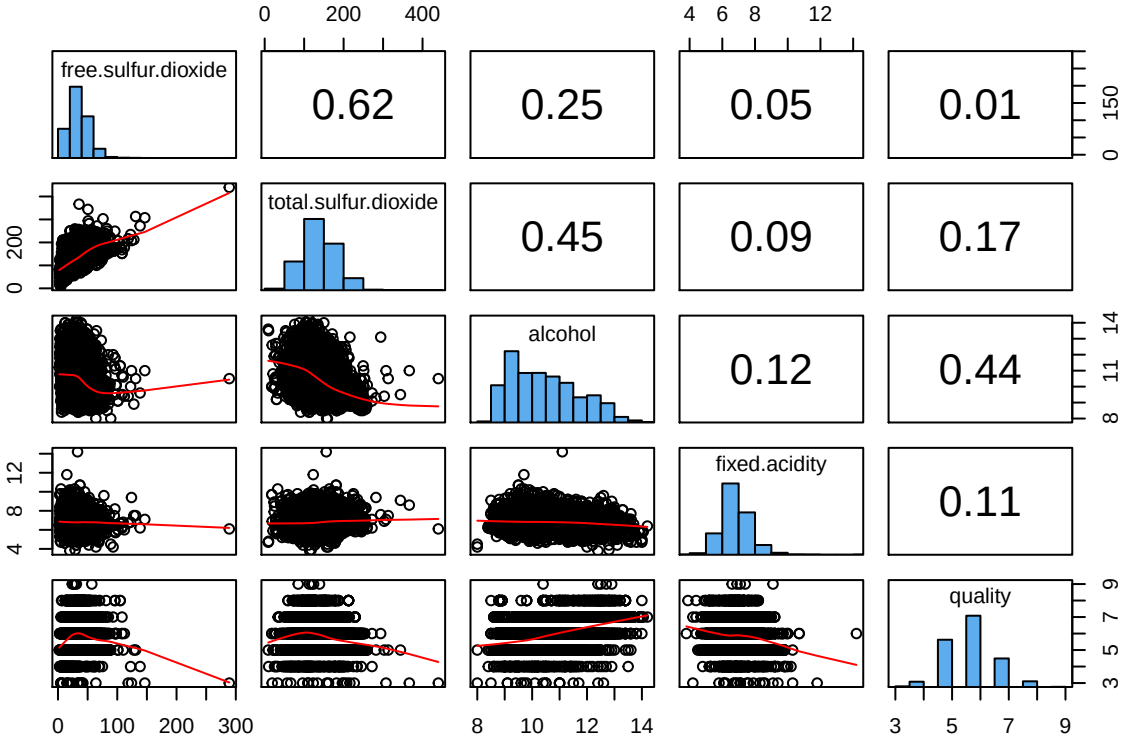


Fig 1: Wine Quality count distribution

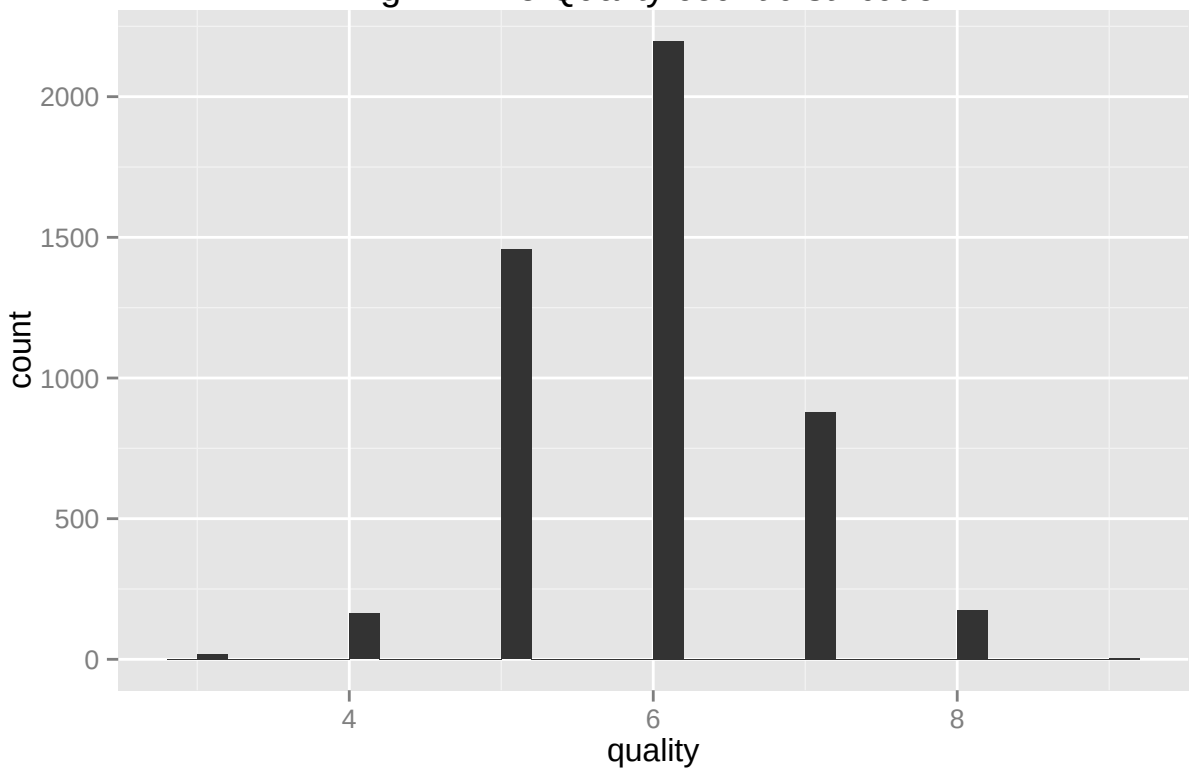
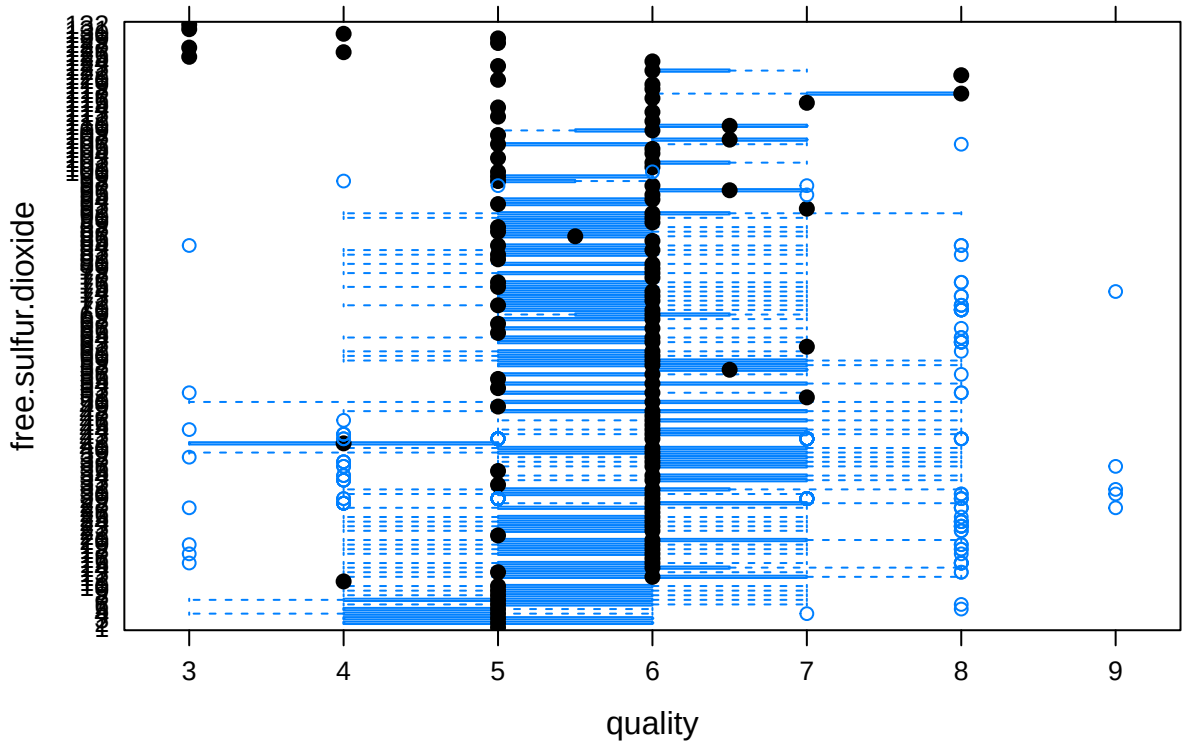


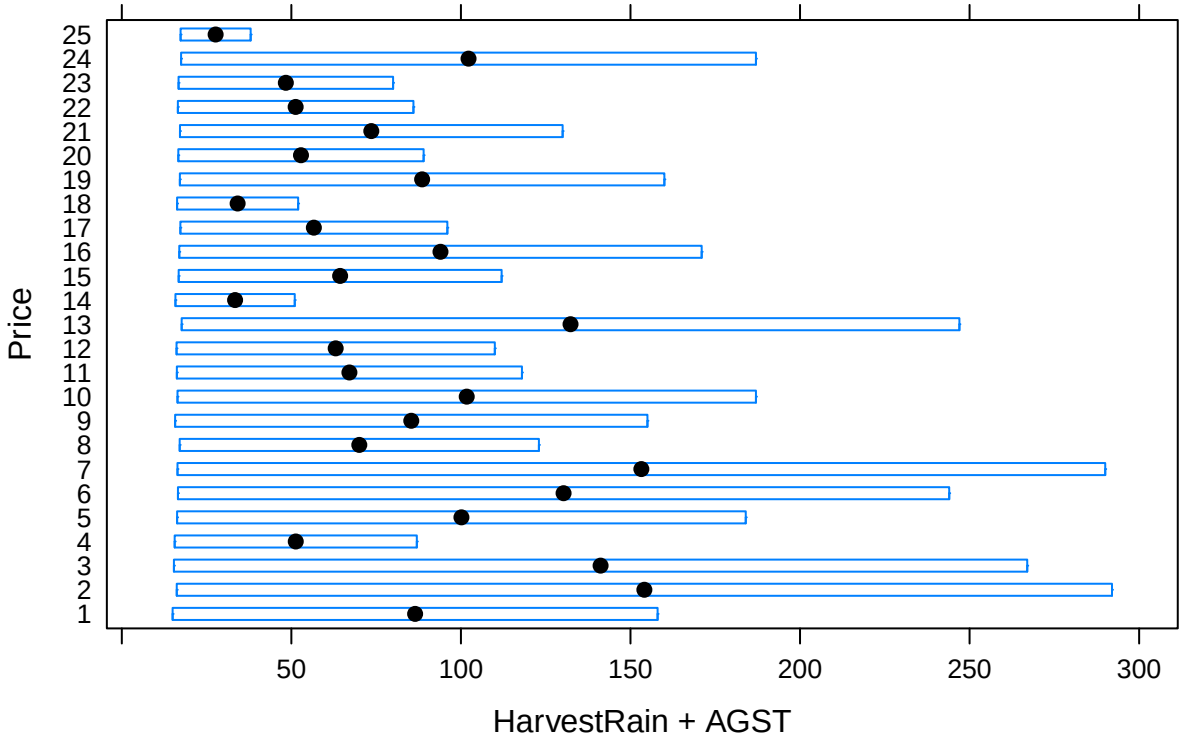
Fig 2: Free Sulphur Dioxide content againts Quality



The plot above shows how free sulphur dioxide content predict the quality of wine to such a huge extent. As can be seen that the quality of the wine follows more or less a normal distribution.

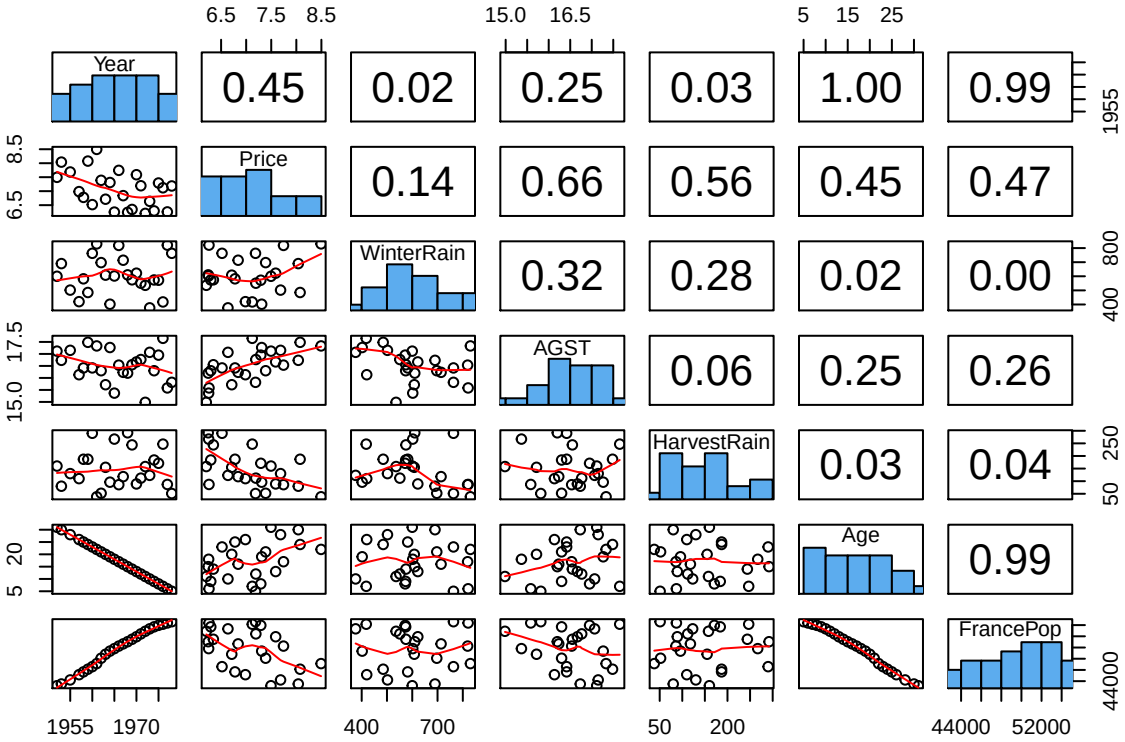
The second dataset is part of the liquid assets dataset which is being used to see if the quality of wine can be better predicted through a regression model than wine testing.

ɟ 3: Price against HarvestRain and Average Growing Season Temperatu



Now also harvest rain and the average growing season temperature matters quite a lot in terms of the price of wine and in order to explore the intricate relationships between all these, various approaches have been proposed in this section.

Now wine prices is the another factor that we can determine from the wine dataset. In order to explore the various properties of the wine dataset here is a diagram of the same that shows the correlation among each of the variables in the wine dataset.

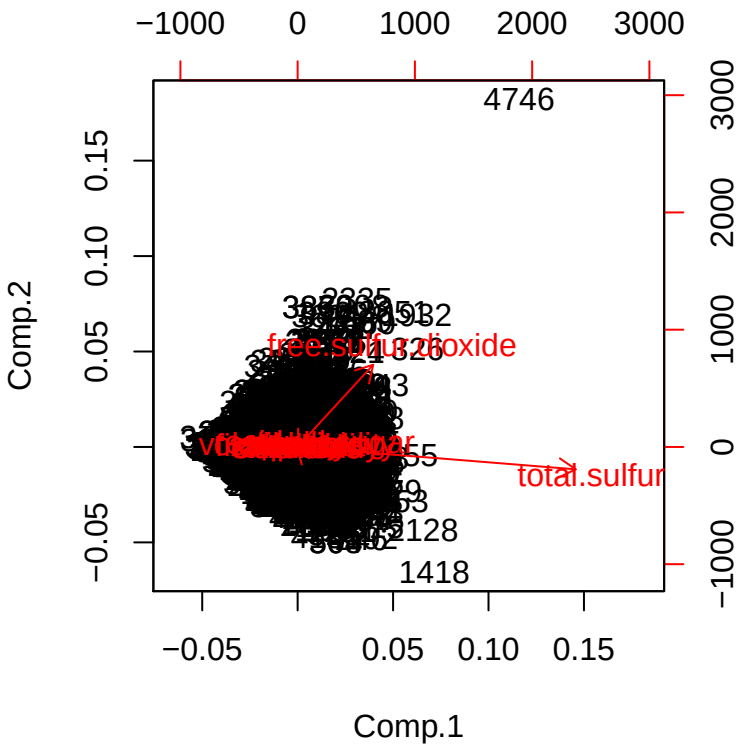


Now somethings to note about the dataset is that some of the correlations are really high but they are intuitive and of not much help to our model.

For instance the population of france and age of the wine. Now since the population has increased with time and age increases with time, this correlation is obvious and of not much help to the model. This is also similar to the correlation with the year variable where the correlation is also significantly high. The fact that the correlation between age and year is one is because they both increase at the same rate, that is yearly in the period in which the data has been collected.

Thus this project is trying to accomplish quite a few things. Can the model perform better than the wine testers and can various attributes that has been obtained through pysiochemical tests be as good as predicting the quality of the model.

The principal component analysis of the whitewine dataset shows that two particular componets account for the most amount of variance. They are as follows. Free Sulphur Dioxide and Total sulphur Dioxide are the two variables that play a very significant role in accounting for the variance.



1.2 Goal:

The output is based on sensory data, while the input is a result of objective, physicochemical tests. The goal is to model wine quality based on physicochemical tests.

This model will also be used to see if the model performs better than the wine testers prediction. Such model is useful to support the oenologist wine tasting evaluations and improve wine production. Furthermore, similar techniques can help in target marketing by modeling consumer tastes from niche markets.

The goal in the second case is to be able to predict the wine cost from the various attributes of wine as mentioned in table 2.

This way not only the wine price can be predicted, these two models can also say which wines will perform the best in the market even in terms of taste.

1.3 Metrics

The measurement of the quality of wine on the test set for good and bad wine on the first wine dataset using generalized linear models. And the measure of prices of wine and various features of the input variables.

- The RMSE on the test set.
- Adjusted R^2 value of the model.
- The AIC of the models that were compared.

1.4 Hypothesis

- The hypothesis is that the quality of wine prediction from the model is not better than the wine tester's prediction.

- From the principal component analysis it could be seen that total sulphur dioxide and free sulphur dioxide account for most of the variance. Thus the hypothesis is to see if sulphur dioxide free wine is better in quality than wines with sulphur dioxide.
- That wine prices depends on the winter rain more than harvest rain at the 95% confidence interval.

2 Approach:

Since all the input variables may not be important in the prediction of wine quality on the test set, we will chose only the most significant ones. Our approach will be to analyze how much each of the dependent variables contributes to the predictive ability of our model. To test the significance of each variable, we will compare values of parameters like measured R^2, adjusted R^2 and sum of square errors. We will also have to look for collinearity and multicollinearity between independent variables so as to reach to a model which is the most accurate in predicting the dependent variable, quality, and at the same time, is simple and more interpretable.

The following plots will show how these variables are related to the quality of wine.

2.1 Data

The dataset contains information from red and wine vinho verde wine samples, from the north of Portugal. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.) from the first dataset. The classes are ordered and not balanced (e.g. there are munch more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant. Thus quite a few feature selection methods, stepwise regression methods, and robust linear regression models have been used to determine the best features. A list of variables in the first dataset is as follows:

Table 1: Whitewine dataset

Attribute Information	
Attribute Number	Variable Name
1	Fixed Acidity
2	Volatile Acidity
3	Citric Acid
4	Residual Sugar
5	Chlorides
6	Free Sulphur Dioxide
7	Total Sulphur Dioxide
8	Density
9	pH
10	Sulphates
11	Alcohol
Output Variables	
12	Quality (score between 0 and 10)

Now because of the distributution of the dataset, the goal esseentially is to predict if the wine quality is good or bad. Wine quality with a quality marked less than equal to 5, it is marked as bad and wine qualities with a quality value of greater than 5 is marked as good wine. Thus a logistic regression seems a really good approach for predicting that.

The second dataset however gives us other features about the wine dataset, each of which has been described below.

Table 2: Wine Dataset

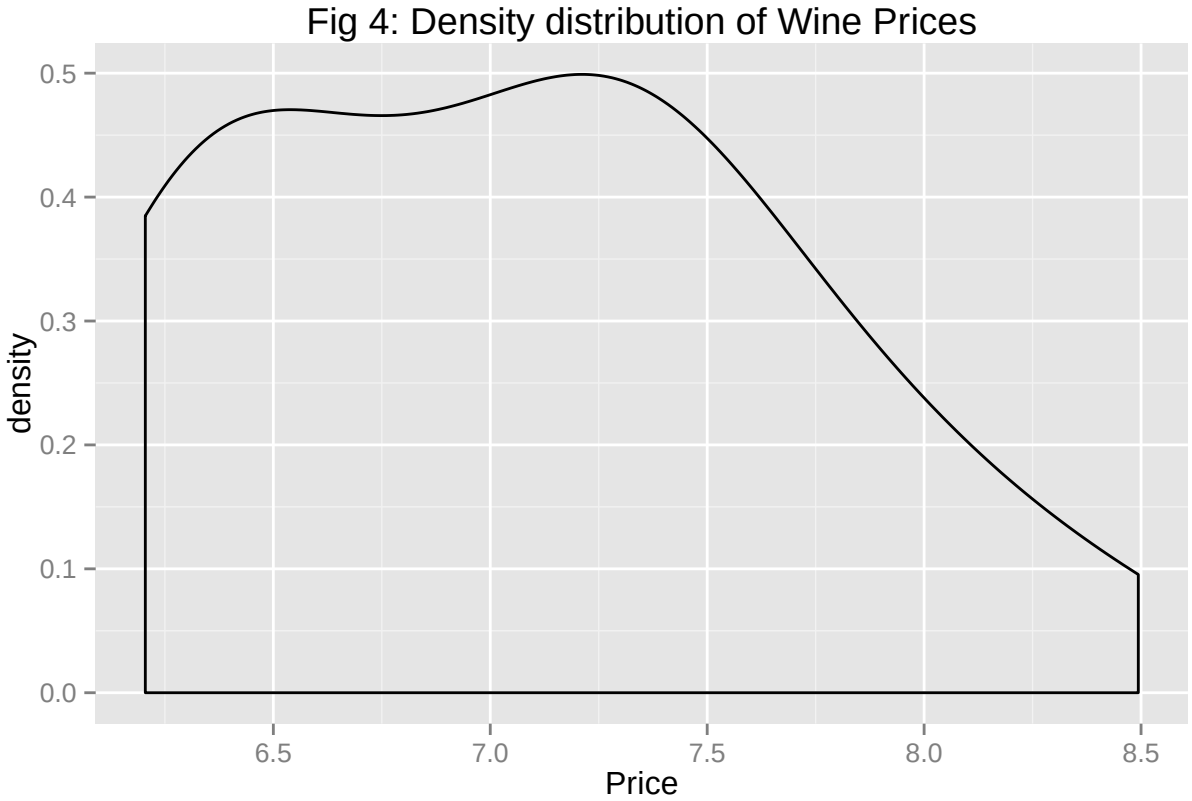
INPUT VARIABLES		
Variable Number	Variable Name	Description
1	TIME_SV	Time since Vintage (Years)
2	VINT	Number of years since the wine was manufactured
3	WRAIN	Winter (Oct.-March) Rain,ML
4	DEGREES	Average Temperature (Deg Cent.) April-Sept.
5	HRAIN	Harvest (August and Sept.) ML
OUTPUT VARIABLES		
1	PRICE	

Summary gives us high level statistical information about the data. The quality is seen to vary only between 3 and 9, with the average value being 5.878. There are no missing values in our dataset.

Now, to start working on the model we need to uniformly split our data set into two sets: training set and test set. It is also important that our test set is a representative of our training set. Sample split randomly splits the data set, but it also makes sure that the outcome variable is well balanced in terms of the predictor variables.

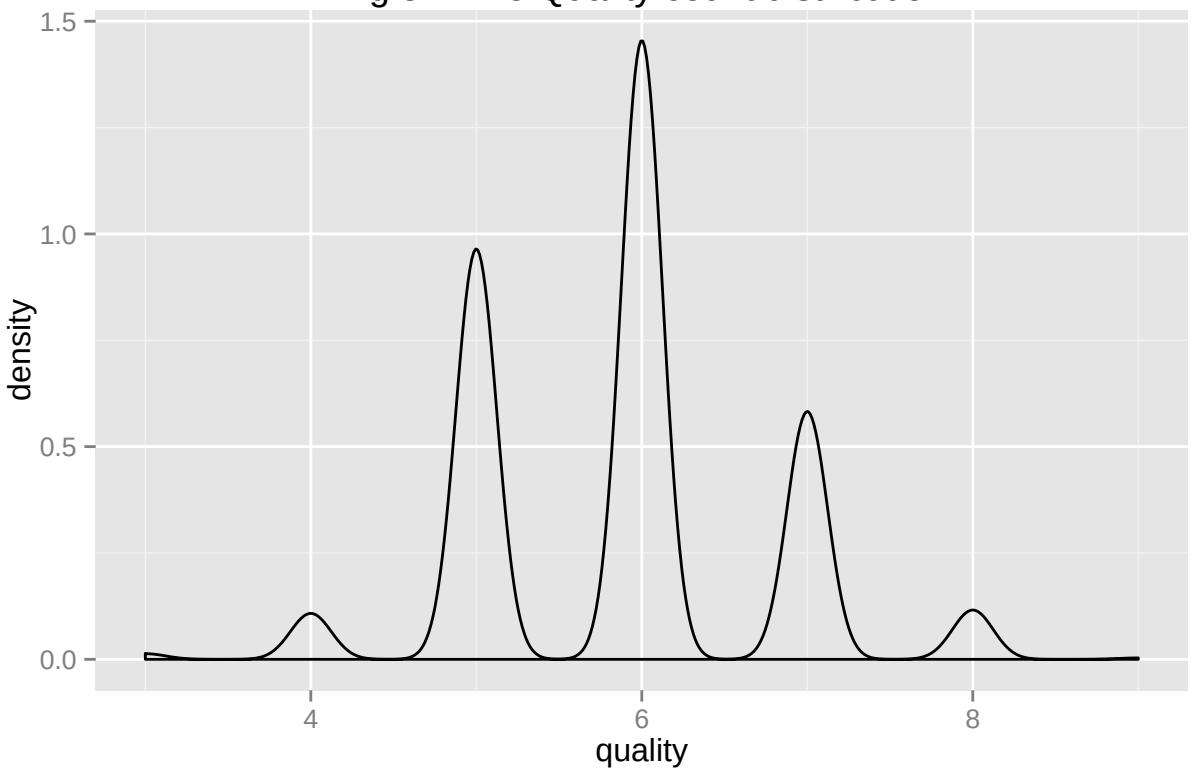
The two datasets does not have any missing values. Thus no correction was necessary for correcting the data. The other biases have been mentioned before in the dataset.

The density of the wine prices is shown below.



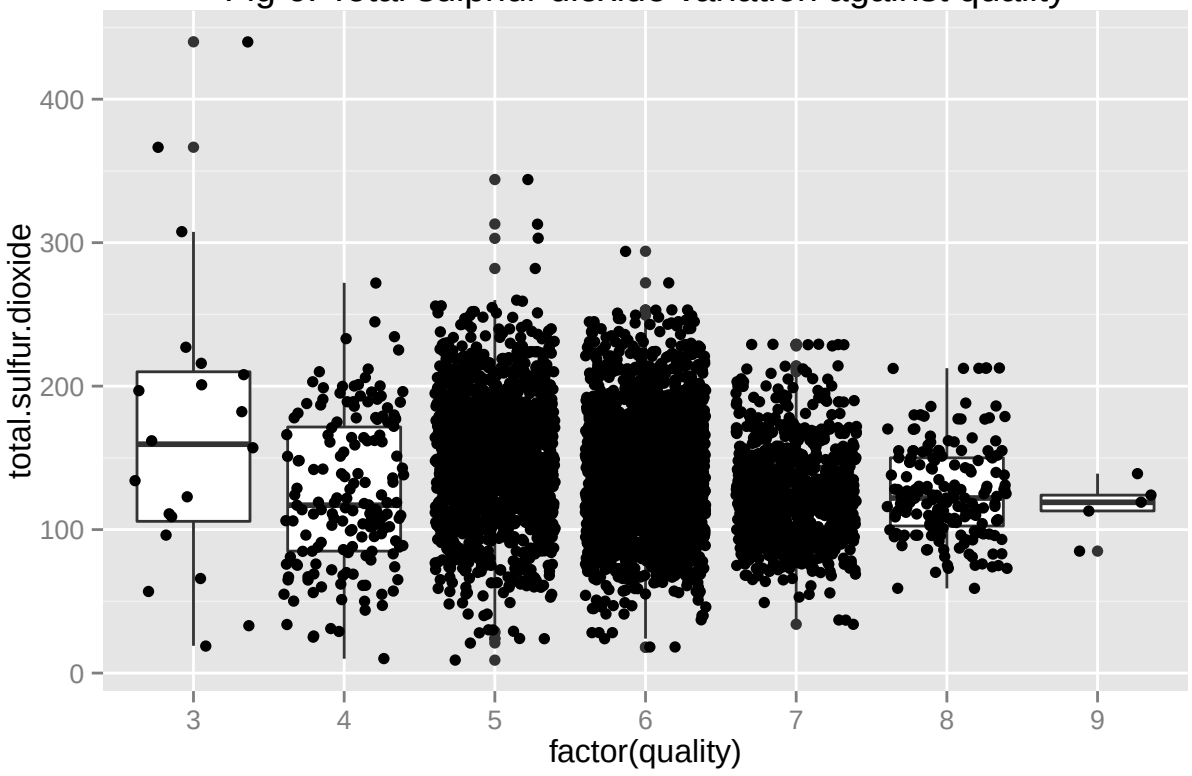
The density of wine quality distribution is shown below. Now the density shows that the quality of wine varies and that some qualities are more in number than the ones. Essentially a lot of mediocre wines are available and very few really bad wines and really good ones are in the dataset.

Fig 5: Wine Quality count distribution



But in order to get a complete picture of how sulfur dioxide content varies according to the quality of wine, the following graoh shows that most wines are mediocre and the quality is between 5 and 7.

Fig 6: Total sulphur dioxide variation against quality



2.2 Analysis

Now for the analysis of quality, it is a categorical variable. It has a few stages as mentioned before. So wines with grades of less than 5 have been assigned a binary value of 0 and wines with a quality greater than 5 binary value of 1.

Now the analysis is two fold, first the quality of wine is being predicted using generalized linear models from the various physiochemical elements and then the wine price is being predicted using robuts linear regression models.

In order that the results are reproducible, a seed is set to a randomly picked number 88. The R library CaTools have been used which essentially divides the data into train and test sets such that the distribution of the predictor variables is more or less equivalent in training and testing.

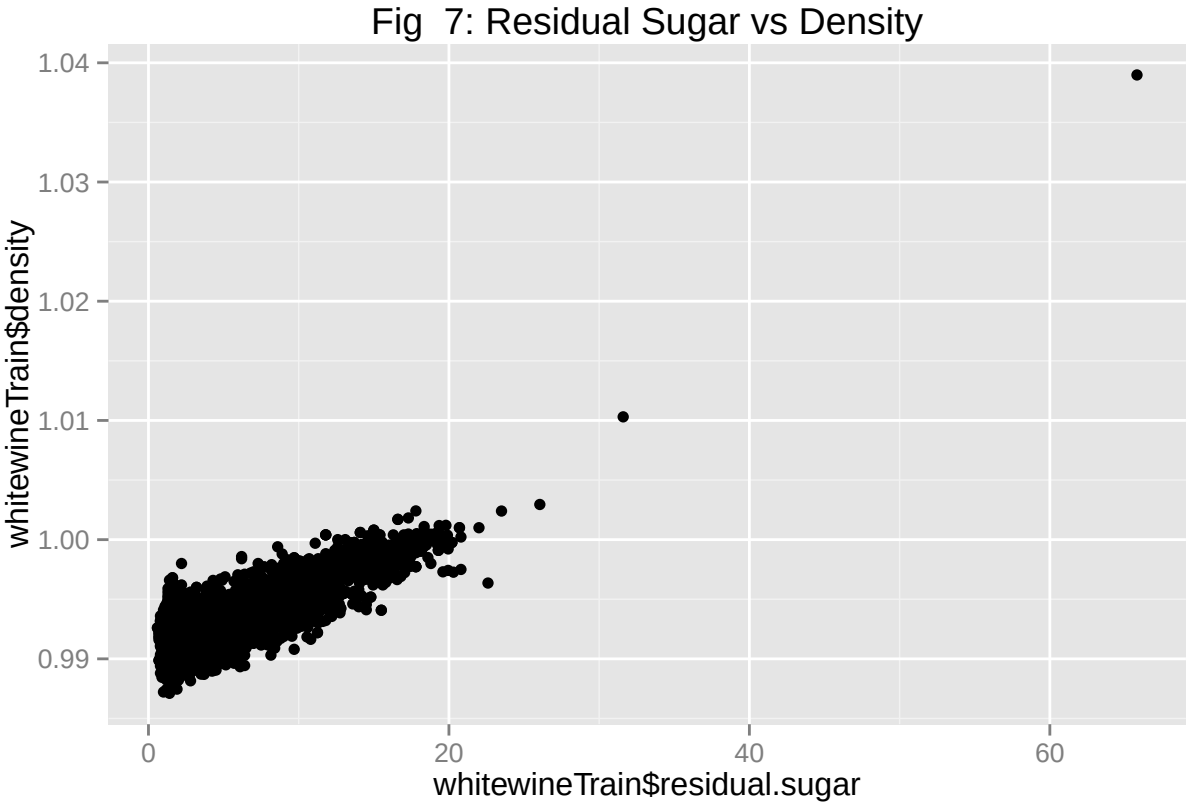
Now a generalized linear model is being constructed and evaluated against the null model where quality bin is being regressed against 1, which evaluates to be the average. This is a good base case since it takes into account the majority and about half the wines in this dataset have a quality less than 5 and the remianing half have a quality greater than one, as can be seen in the density graph above.

Now this can also be verified from the diagnostic plots which have been shown in the evidence section.

As can be seen that the model with the various physiochemical test results and their properties is significantly better than the null model to which it is compared to for the quality.

Variables are removed separately and different combinations and selected different combinations to compare different models in terms of better R^2 values, sum of squared errors and simplicity. Collinearity and multicollinearity is also checked for in the model, and some variables were reduced to prevent overfittingas explained below.

Residual sugar and density show a high positive value of correlation(= 0.83978900).



Multiple R-squared: 0.2572, Adjusted R-squared: 0.2552, SSE: 2140.143

As expected, it is obviously better to remove only density out of the two.

We know that mearsured R^2 always increases with increasing number of variables in the model. So it is expected that model1a or model1b will have smaller measure R^2 compared to model1 which has all variables. We would have preferred a greater adjusted R^2 though. This is because adjusted R^2 decreases when there are insignificant variables in the model. From the fact, that residual sugar and density are highly correlated is it reasonable to remove one of them from to model to make it simpler.

The final model we selected is thus a model on which stepwise regression was done on a generalized linear model, where the logistic regression was done on the categorical variable quality against all the other predictors.

The other dataset has numerical variables as the predictor variable and a robust linear regression or an rlm model is being used in this case.

But the above model takes care of all the variables but some variables make more sense than some others.

A stepwise regression is being used here to find a solution to the problem.

The other part of this project concentrates on the prediction of wine quality price from various natural properties as mentioned in the other table before.

For this, price is being predicted from the various qualities of wine.

Now before starting the prediction model, this data is checked for any serial correlation and the auto correlation and the partial auto correlation plots are plotted to get a detailed over view of the data.

Fig 8: ACF plot of prices

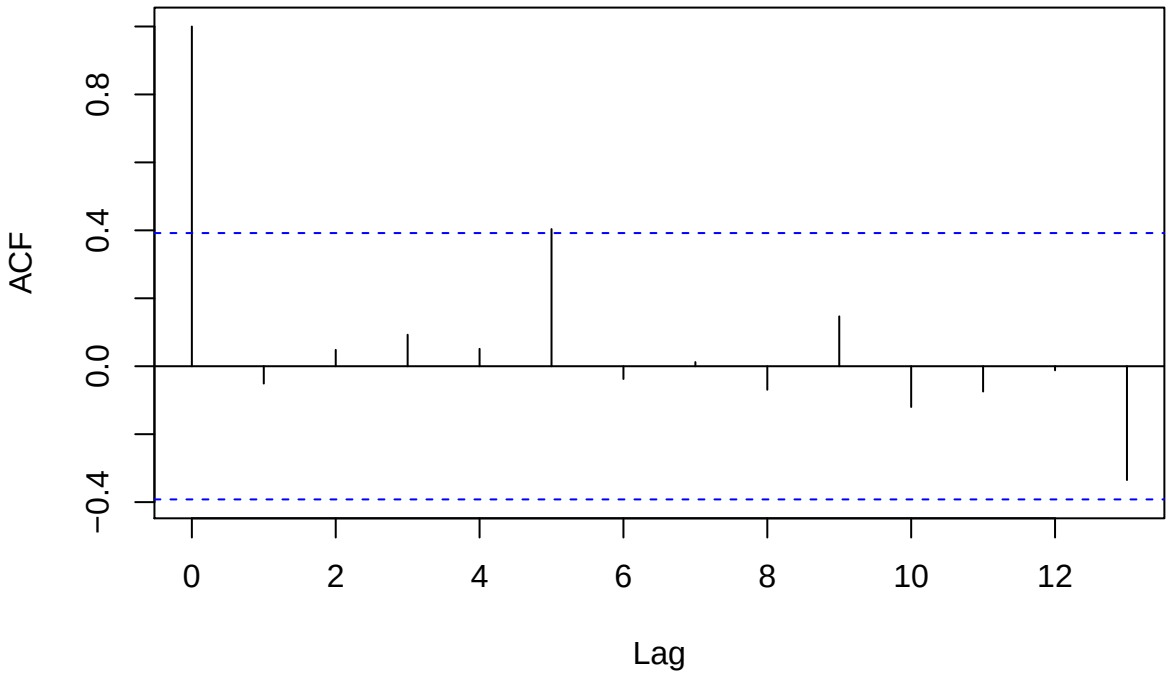
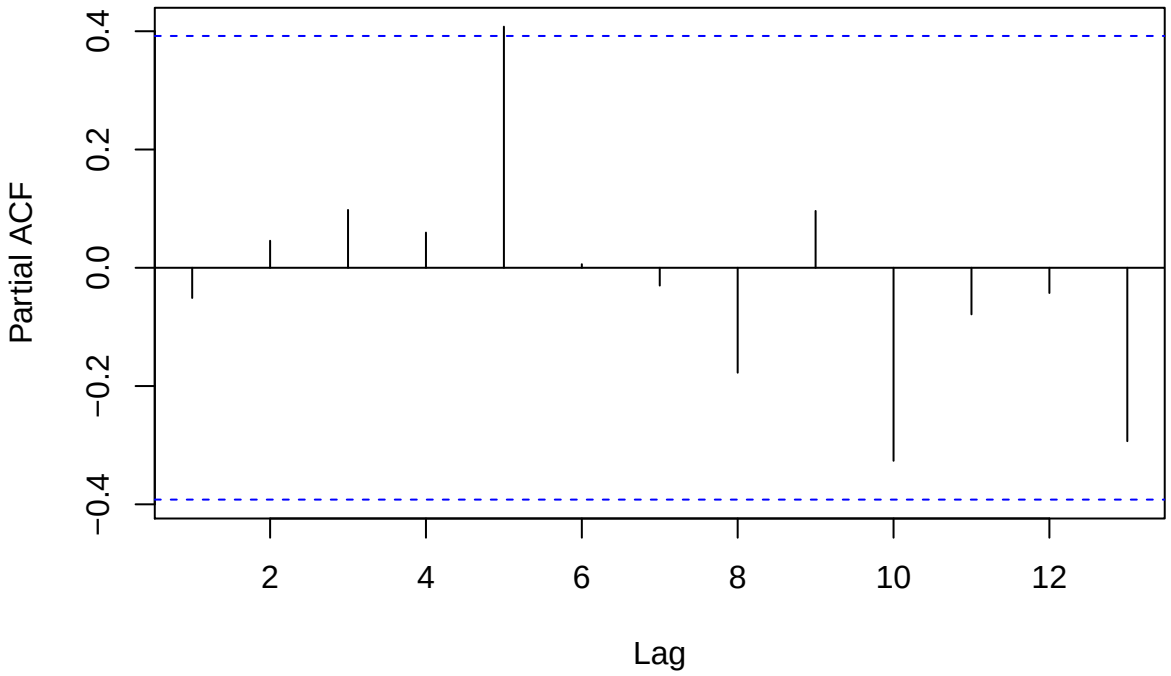


Fig 9: PACF plot of prices



As can be seen from the plots above the data does not have any serial correlation and thus a time series analysis in this case is not necessary.

Now as can be seen the average growing season tempearature and harvest rain play a very important role in the wine quality prediction and winter rain plays the next most signifcant role.

However in order to assess the model better, a stepwise regression method is applied on this model so that the most significant features are found. The best model is being selected on the basis of the Aikaike Information Criterion.

The final model has the parameters, WinterRain, Harvest Rain and average growing season temeparature. These play the most significant role in predicting the price of the wine.

Now a test set is being used to evaluate the quality of the prediction and analyzing the performance of the model.

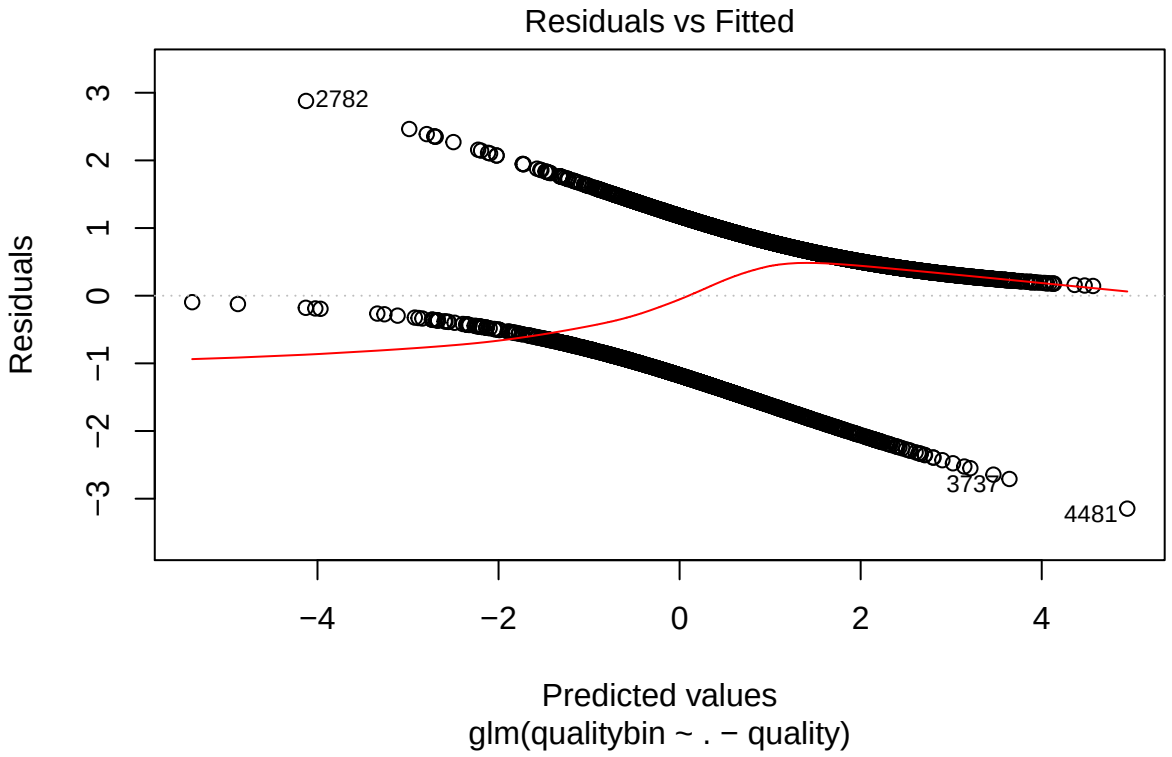
This model is then compared for R^2 and adjusted R^2 to evaluate the quality of the model and it can be seen that the sum of squared errors on the test set is 0.2626265, which is significantly better than the previously mean squared error of 0.7566484. Also the partial F-test, since this is a nested model shows from the p-value that the second model is significantly better than the first base model as was mentioned.

3 Evidence

Now the following diagnostic plots show how this model is performing. Also the distinct differences between the distinct sections show the two separate sections for good and bad wine quality.

Now the residuals vs plotted values show the distinct differences in two sections since the really high concentration of points show good and bad wine respectively.

Fig 10: Residuals vs Plotted values



Now the normal qqplot and the scale location plots also shows a similar trend of two distinct categories of points. The cook's distance plot shows some of the influential points and the residuals vs leverage plot also highlights the two distinct categories. All these show that the model can actually perform really better than this.

Thus an approach similar to stepwise regression in case of generalized linear models is used which helps in giving a final better model.

Fig 11: Normal QQ Plot

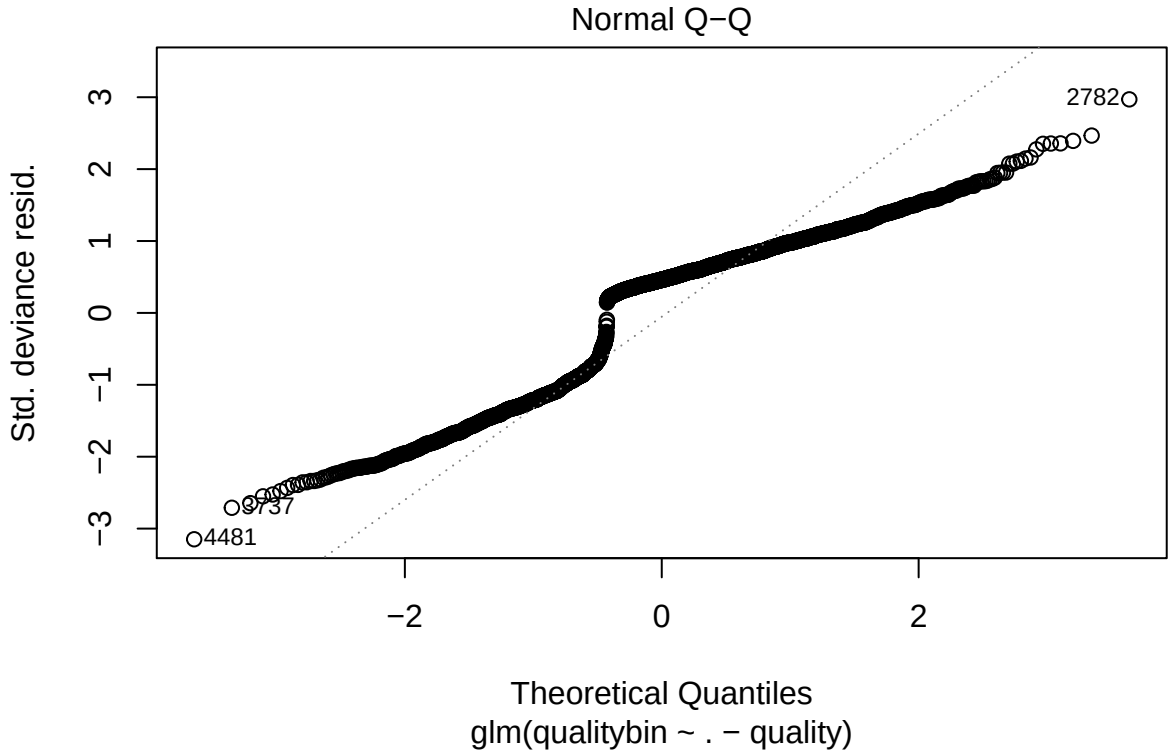


Fig 12: Scale Location Plot

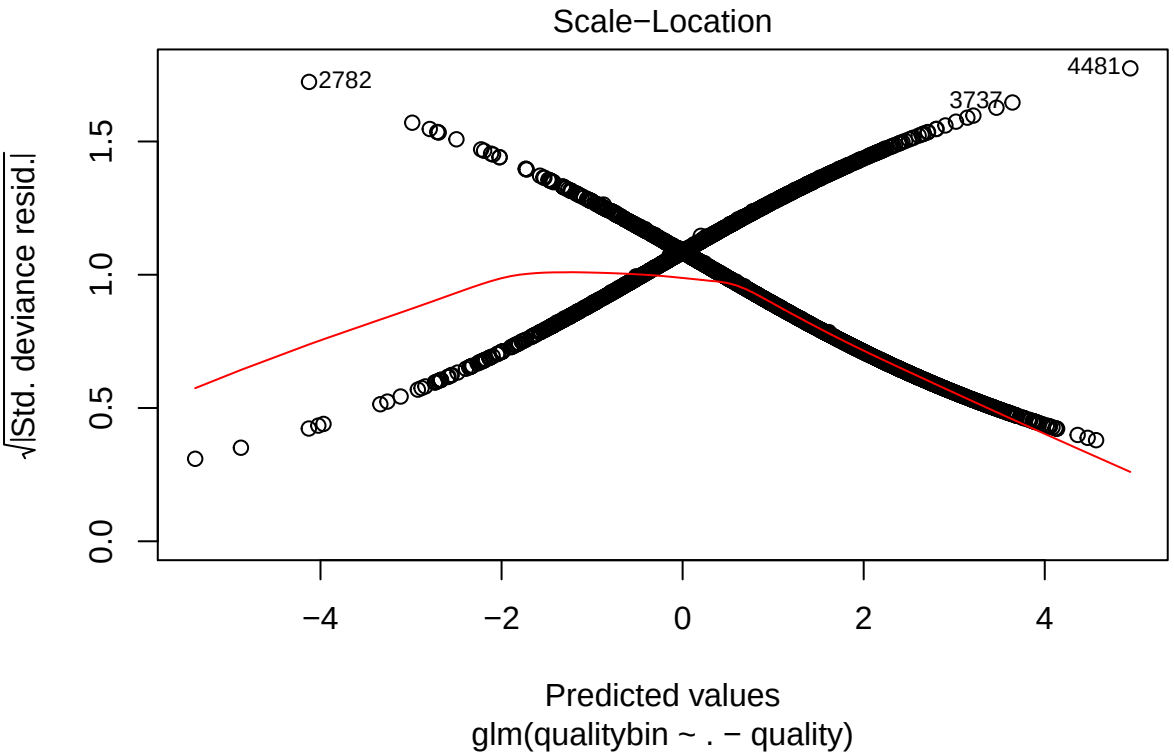


Fig 13: Cook's Distance Plot

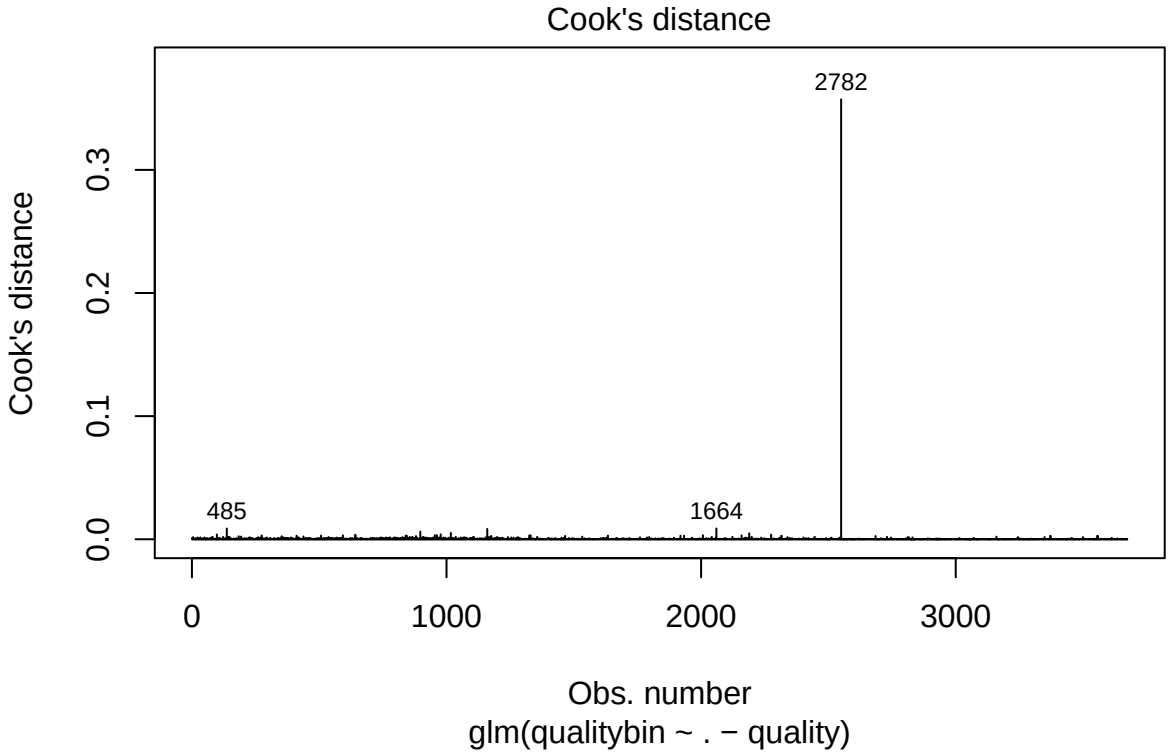
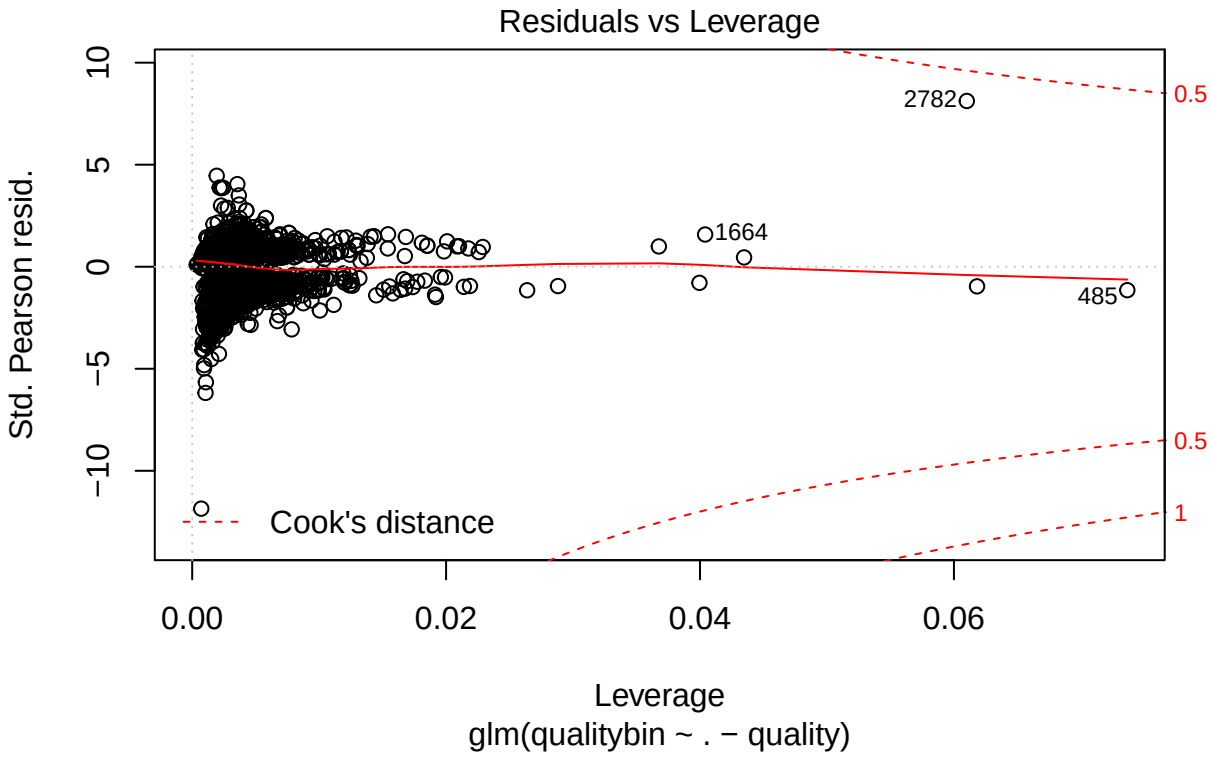
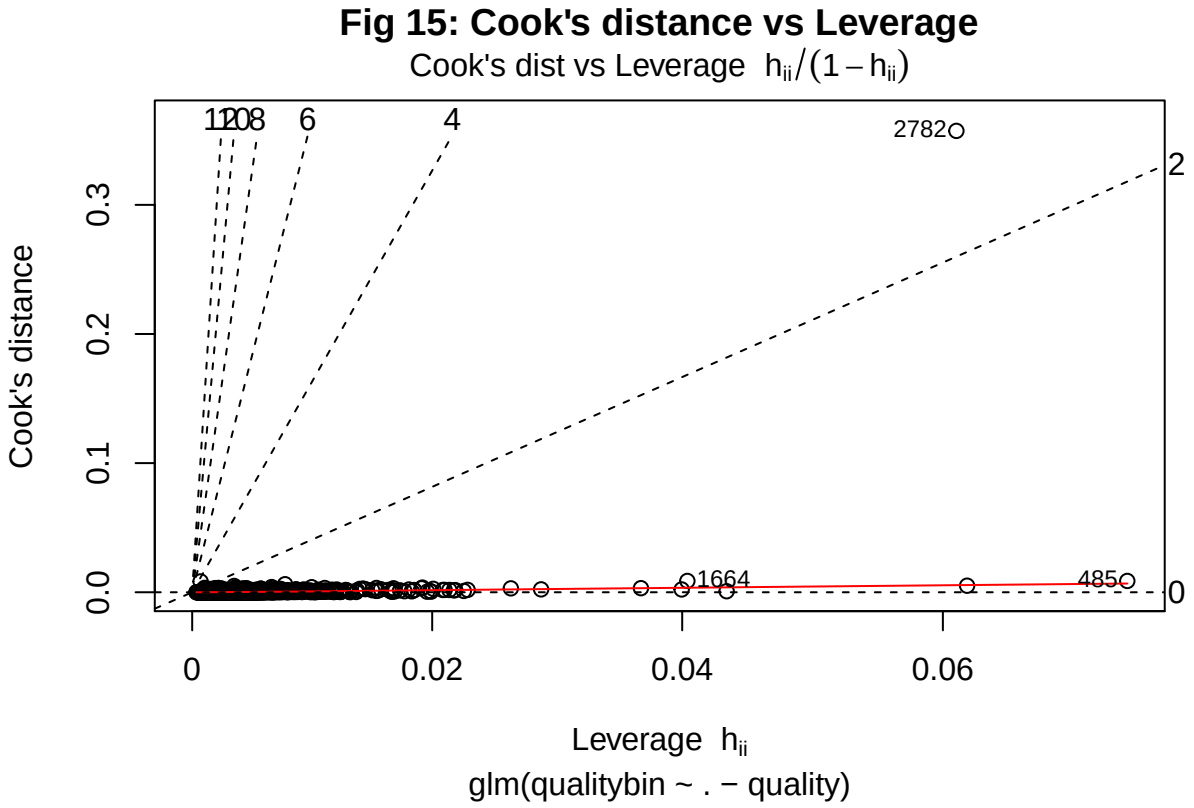


Fig 14: Residuals vs Leverage





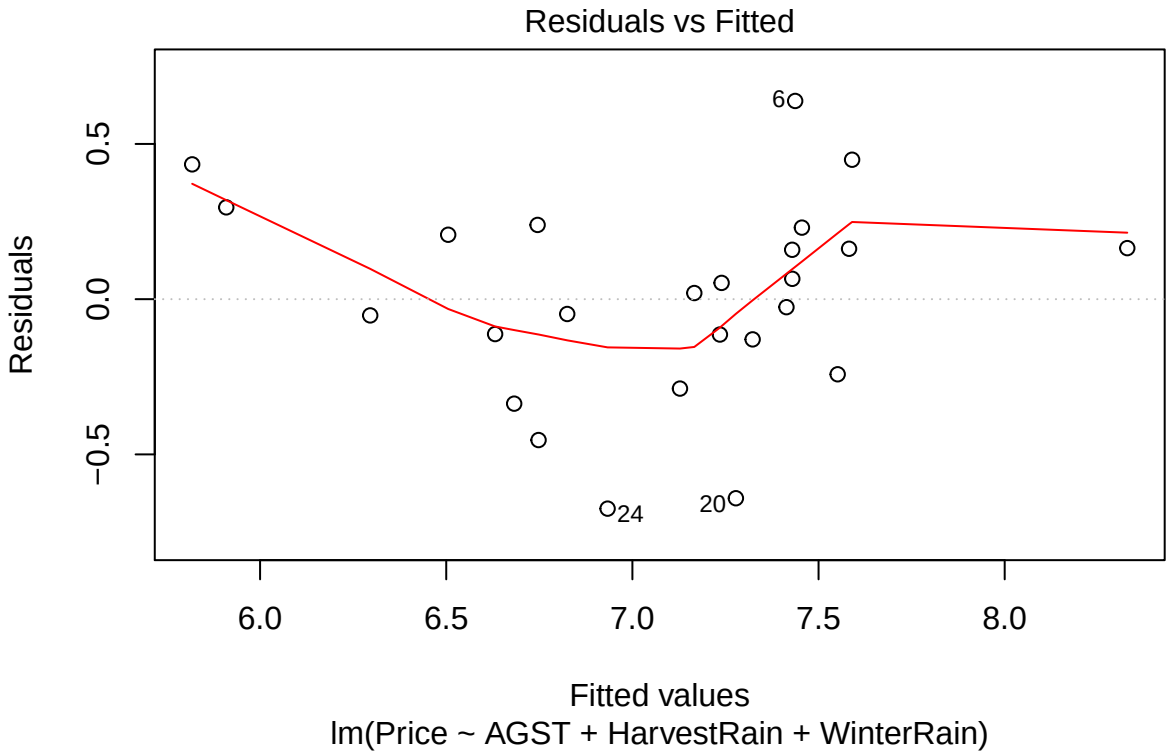
Since this model is doing quite well, a test set prediction is used to evaluate the quality of the model.

The AIC for this model is 26 which is quite good and significantly better than the previous AIC value of 4186 which shows that a wine quality can be significantly well predicted through the exact compositions of various physiochemical tests.

The test set predictions are quite good and it can be easily seen that the model is performing really well. And as it turns out free sulfur dioxide content is really good for the prediction of wine quality.

Now for the wine price prediction, the diagnostic plots have been analyzed below.

Fig 16: Residuals vs Fitted



The residuals vs fitted plot performs quite well and the mean of the residuals is 5.63243e-19, which is almost close to zero thus showing a very good fit.

The other diagnostic plots are below. The normal QQ plot shows a very good, almost normal distribution.

Fig 17: Normal QQ Plot

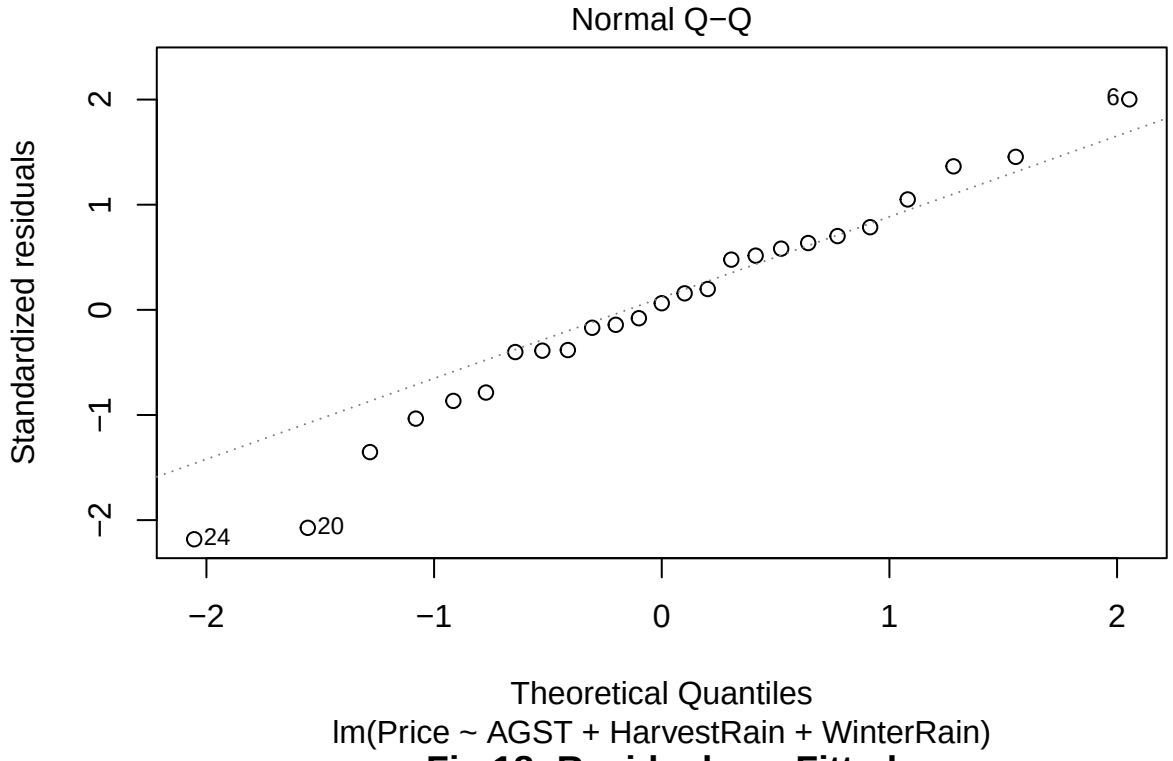


Fig 18: Residuals vs Fitted

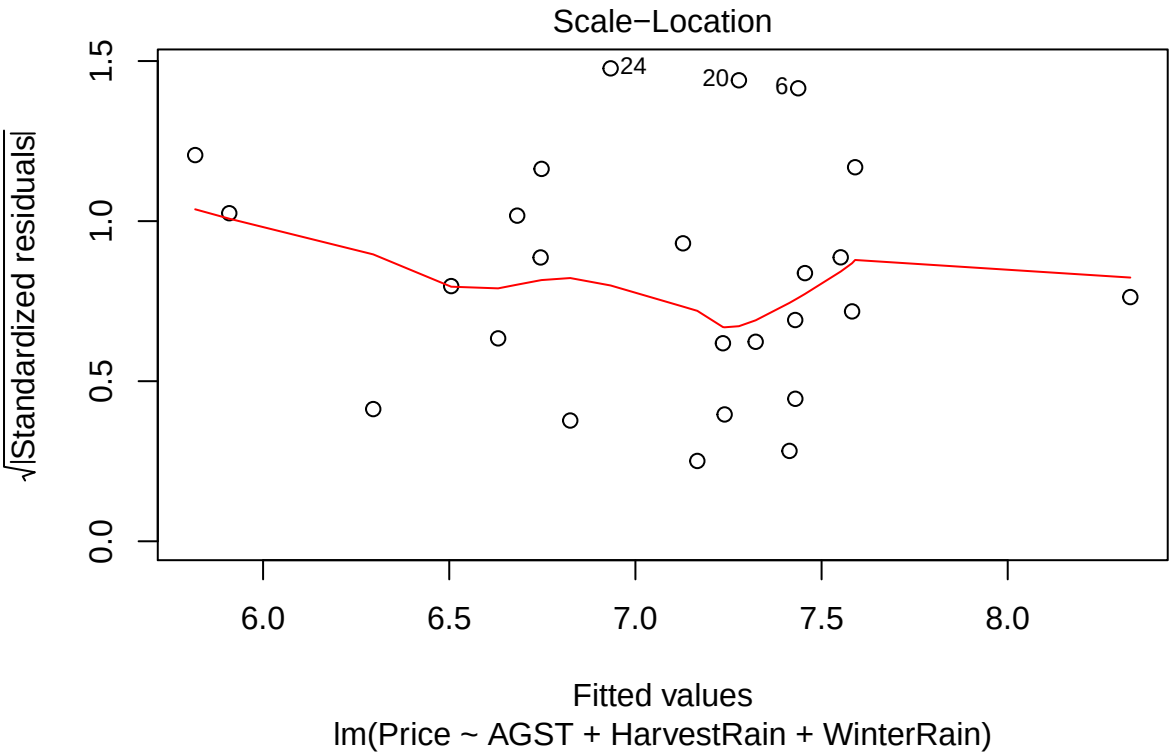


Fig 19: Cook's Distance

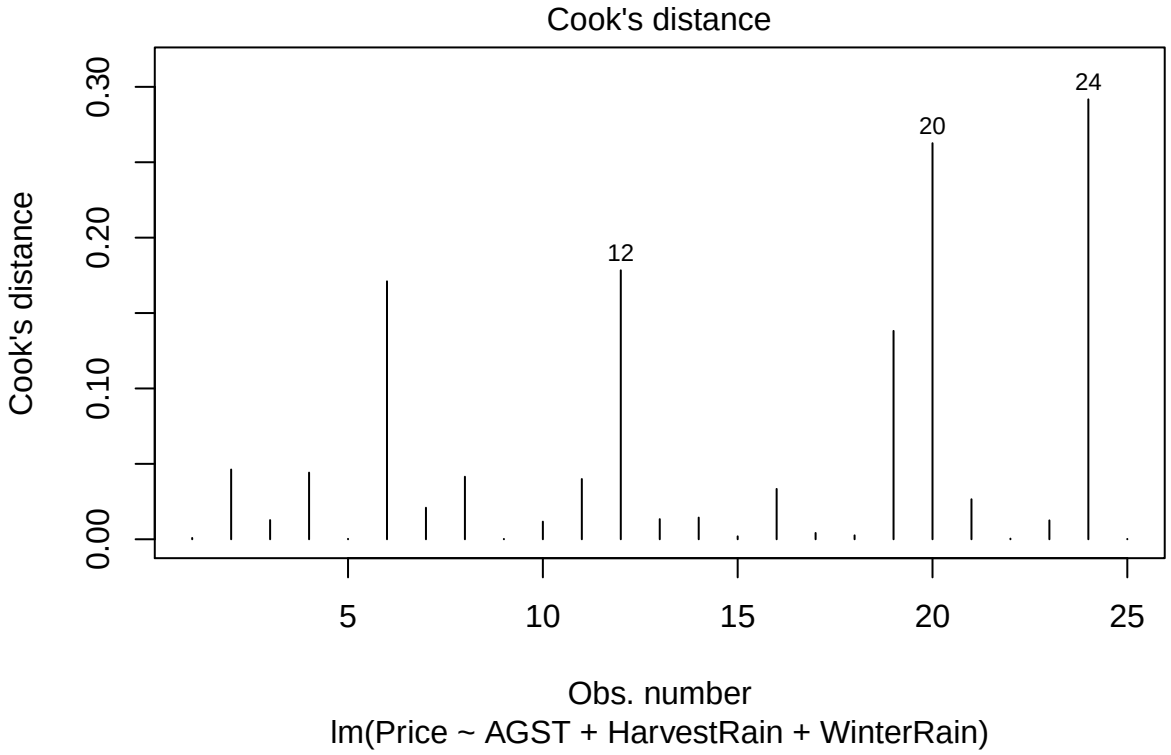
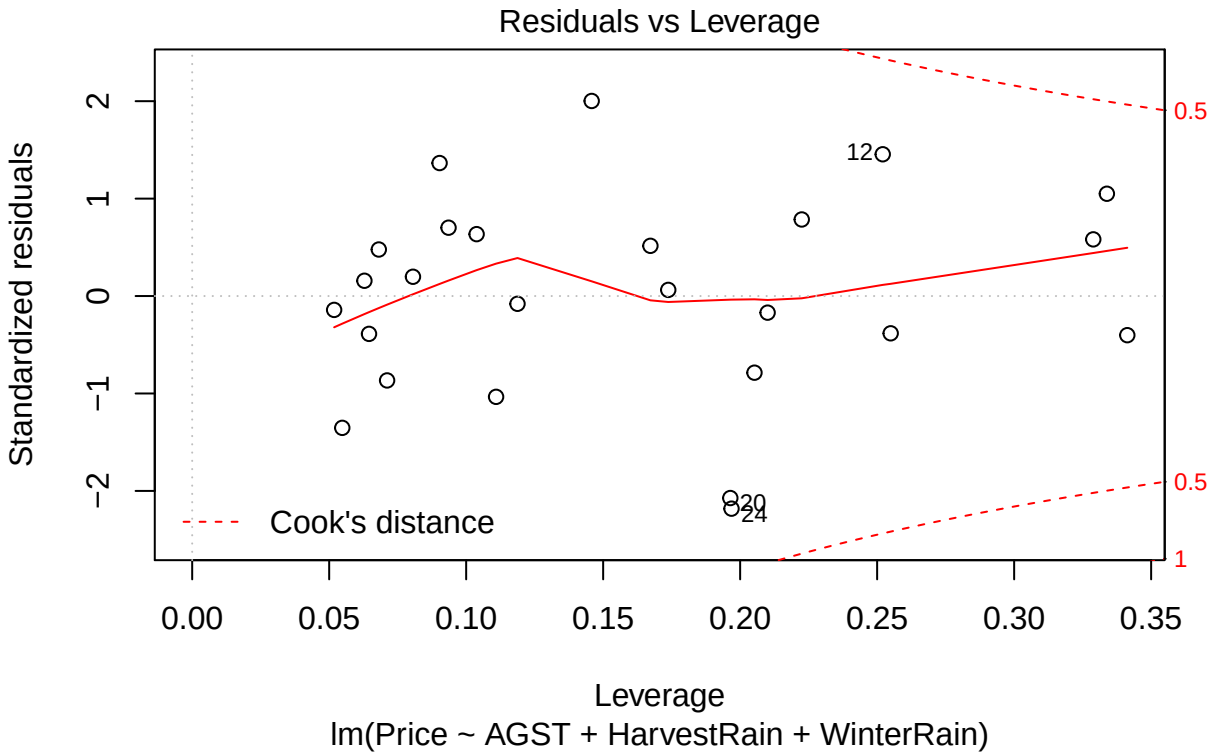
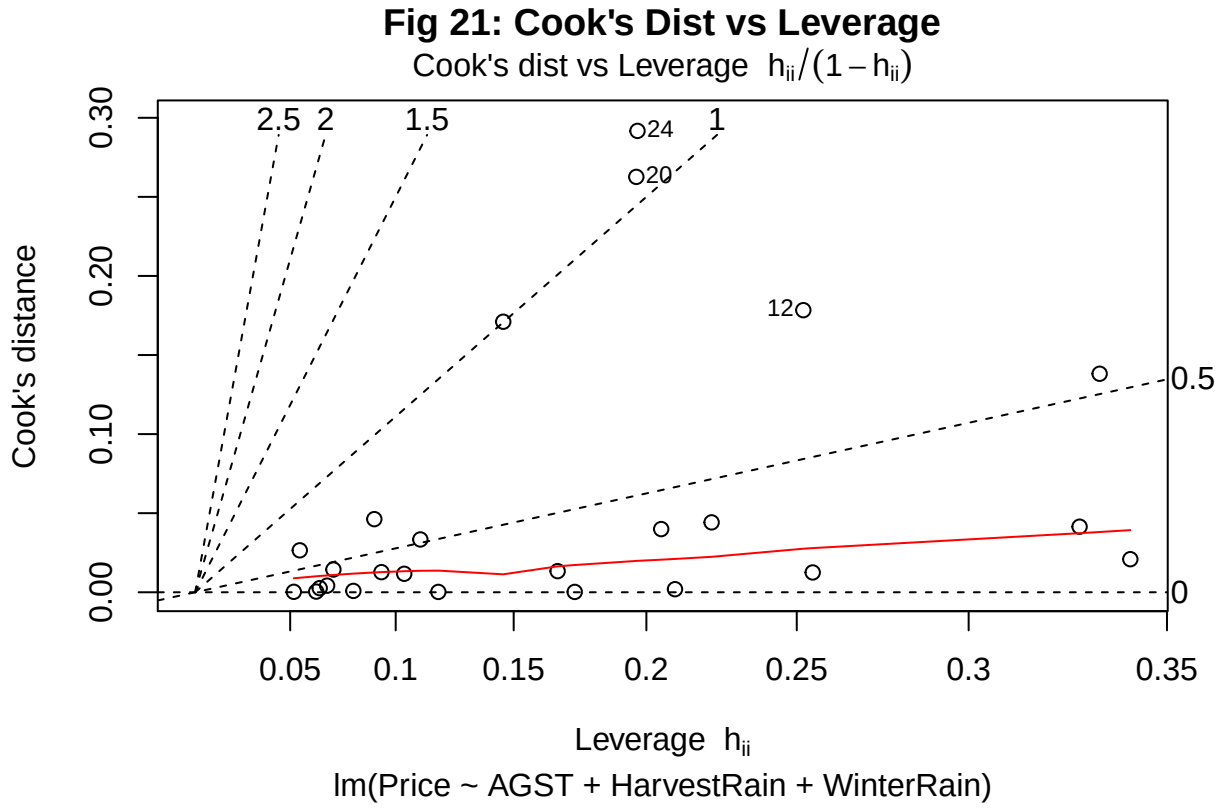
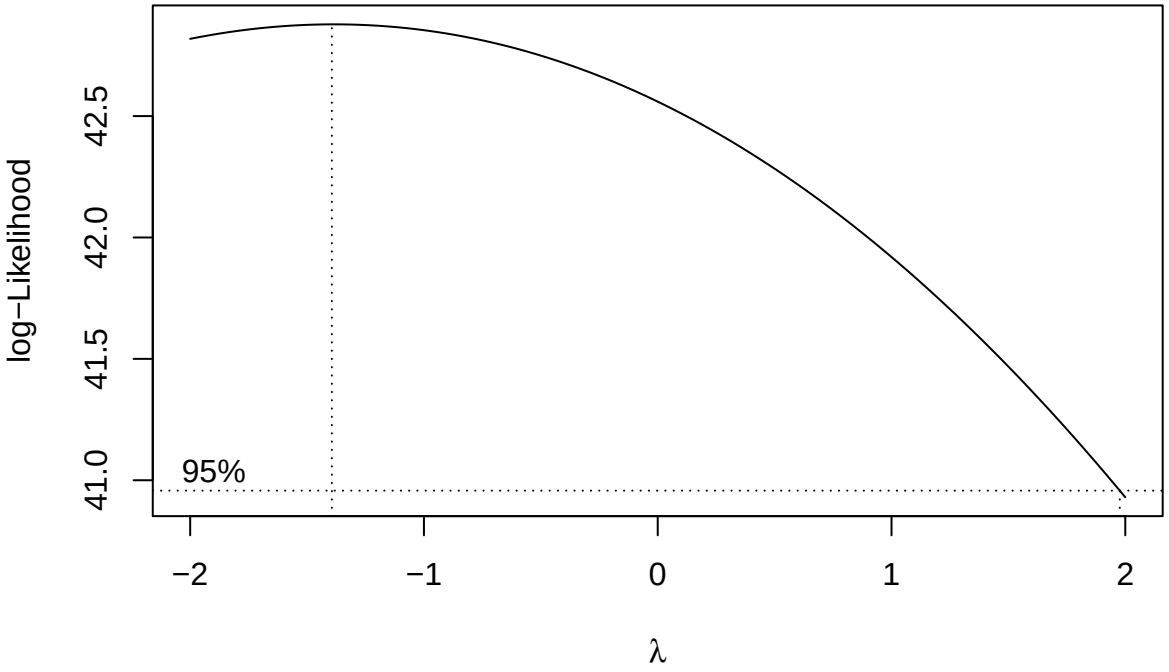


Fig 20: Residuals vs Leverage





The boxcox shows a fairly good fit and thus no further modification was necessary for the improvement of the model.



Now bootstrapping was used for better prediction of the regression models and the following results were obtained.

The wine quality test dataset performs quite well. The prediction accuracy of this model is about 0.84. The precision also happens to be 0.84.

The bootstrap statistics are as follows:

Table 3: Bootsrap statistics of the final model

Bootstrap Statistics		
Original	Bias	Std. Error
-4.3016	-2.97844e-01	2.3738
0.68102	1.76319e-02	0.13481
-0.0039	-8.812e-06	0.00095
0.00176	1.10229e-05	0.0006

Table 4: Wine Quality Prediction

Confusion Matrix		
	True (Predicted)	False (Predicted)
True (Actual)	485	89
False (Actual)	104	546

4. Recommendation

As can be seen from the above tests that the preliminary wine quality test can be very easily done using the regression models and the cost of hiring wine testers at the preliminary stage is not necessary. In fact the precision of the model is really high at 0.84 and so is the accuracy. Thus I believe preliminary wine tasting at breweries can be done using machine learning models like these because models like these can take into account various components of detecting the quality of wine from its chemical content. This model performs almost as good as the wine tasters but is not significantly better than them. This means that the cost of hiring wine tasters for every stage is not necessary and breweries can save a lot of cost by not hiring wine tasters at every stage and the preliminary quality testing can be done by breweries through machine learning models and this can indeed save a lot of cost.

The second model about the prediction of prices is fairly accurate and the p-value of the model significance 1.359e-06, and the prediction price accuracy is about 0.74. This shows that the model price can be predicted fairly accurately from natural factors and as it turns out from the model that the average growing season temperature and harvest rain matter significantly in terms of the wine prices than winter rain, which also happens to be a very significant factor. The test set price prediction is quite good and the root mean squared error on the test set is 0.08199167 which shows that the price depends significantly on these factors.

References

[1] L. E. Barnes and D. E. Brown, *Project 1,2 and 3: Code given in class for class projects in SYS 6021, 2014.*

[2] Project 3 template, Project 2 template, Project 1 template, Class template in SYS 6021, 2014.

[3] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

[4] Vinho Verde and informations related to it. *http://en.wikipedia.org/wiki/Vinho_Verde*

[5] Export of Vinho Verde and quality related informations *http://theportugalnews.com/news/first-half-exports-of-vinho-verde-up-14-on-year-germany-overtakes-us/32625*

[6] The wave of Vinho Verde *http://www.thedrinksbusiness.com/2014/03/new-wave-vinho-verde-threatens-regions-landscape/*

A Optional Appendix

Table 5: Final Model Coefficient list 1

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	273.8785	80.1895	3.42	0.0006
volatile.acidity	-6.0686	0.4595	-13.21	0.0000
residual.sugar	0.1813	0.0305	5.95	0.0000
free.sulfur.dioxide	0.0090	0.0031	2.87	0.0042
density	-287.0253	81.2637	-3.53	0.0004
pH	1.2748	0.4069	3.13	0.0017
sulphates	1.7628	0.4135	4.26	0.0000

Table 6: Final Model Coefficient list 2

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4.3016	2.0367	-2.11	0.0468
AGST	0.6810	0.1117	6.10	0.0000
HarvestRain	-0.0039	0.0010	-3.95	0.0007
WinterRain	0.0012	0.0006	1.99	0.0601