Project Report for MATH 261A

Wine Quality Prediction

Shayna Gaulden and Tien Nguyen
San Jose State University MS Students

Report Prepared for Math 261A
Regression Theory
Professor: Dr. Tortora

Department of Mathematics and Statistics

November 2021

Contents

1 Introduction

This project seeks to create a regression model to predict the wine quality of red variants of the Portuguese "Vinho Verde" wine from a data set including 11 variables that can be used as regressors and 1 variable that rates the quality of the wine to be used as the predictor variable [1]. The quality of wine is rated on a scale that goes from 3-7. The regressor variables that are available include all of the following: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol.

2 Methodology

2.1 Preparing the Data

The data has 1599 data points for each variable which seemed to be a sufficient enough amount to safely split the data into a training set and a testing set. All the model fitting was done exclusively on the data split into the training set while at the end two final models were chosen to be tested with the testing set of data. The data was split using a 20% testing and 80% training rule so in the end 1279 data points from each variable were used for training and the rest reserved for testing.

2.2 Simple Linear Regression Model

The first model was fit to all available variables using the following naming conventions.

Y = Quality

X_1 = Fixed Acidity	X_6 = Free Sulfur Dioxide
X_2 = Volatile Acidity	X_7 = Density
X_3 = Citric Acidity	X_8 = pH
X_4 = Residual Sugar	X_9 = Sulphates
X_5 = Chlorides	X_{10} = Alcohol

Simple Linear Regression Model

$$Y = X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 + X_9 + X_{10} + \epsilon$$

Fitted Simple Linear Regression Model

$$\begin{aligned} Y =& 36.89 + 0.06X_1 - 1.20X_2 - 0.34X_3 + 0.02X_4 - 1.59X_5 - 0.002X_6 \\ & - 33.96X_7 - 0.23X_8 + 0.85X_9 + 0.30X_{10} \end{aligned}$$

2.2.1 Model Adequacy

Right away there are potential problems with the estimated coefficients. The range of ratings for quality of wine in our data go from 3 to 8 on what we assume to be a 1 to 10 scale, but our intercept is 36.89 and only one of our variables has an estimate coefficient of 33.96 which might cause some unpredictable behavior. We looked at our model summary and ANOVA table to find the R^2, R^2_{adj}, and MSRES to determine if the model is a good fit for the data.

R^2	R^2_{adj}	MSRES
0.3736	0.3687	0.415

The R^2 value shows that approximately 37% of the variation in the data can be explained by the regression model. This number is not very high. To learn more about this model the underlying assumptions of a simple linear regression model needed to be checked. To do this we looked at the model's residuals.

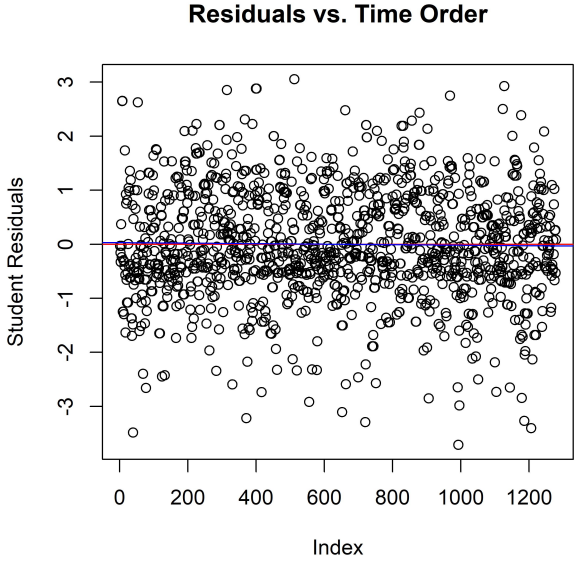

Figure 1: Student residuals plot.

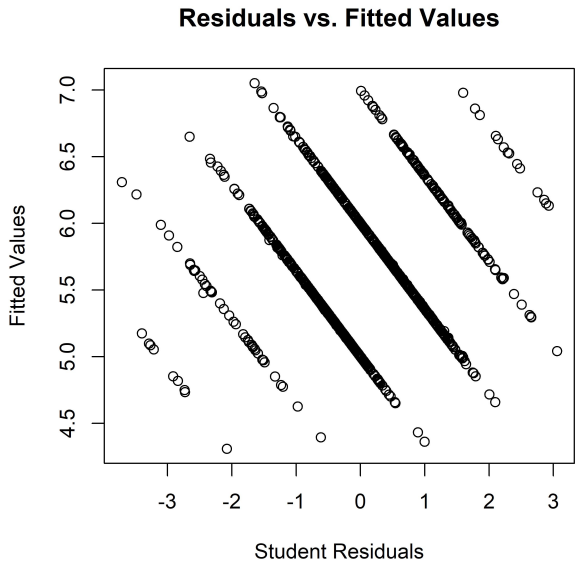

Figure 2: Student residuals vs. fitted values.

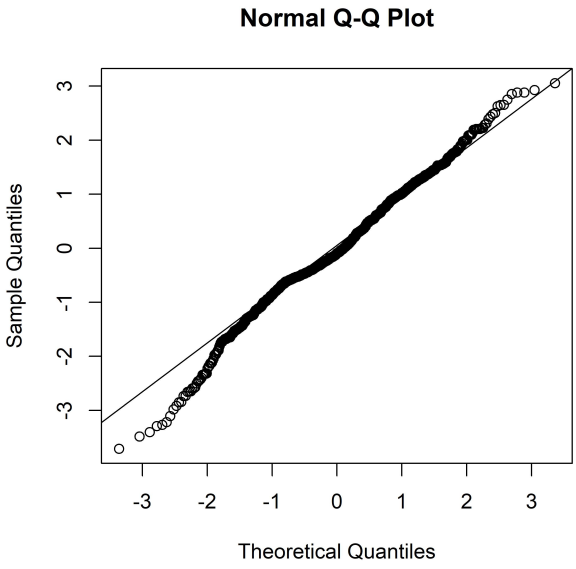

Figure 3: QQ-plot.

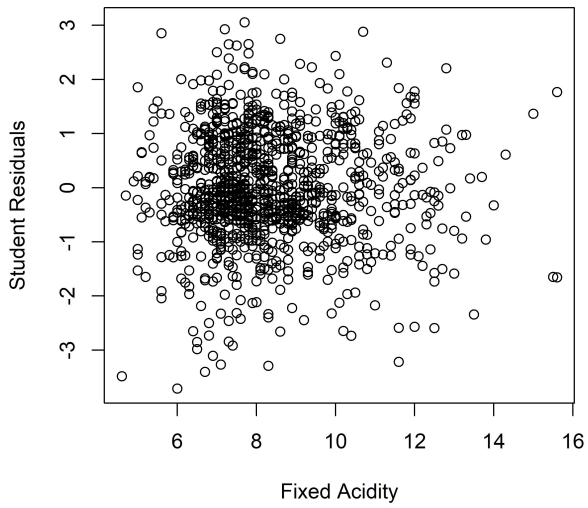

Figure 4: Student residuals vs. X_1.

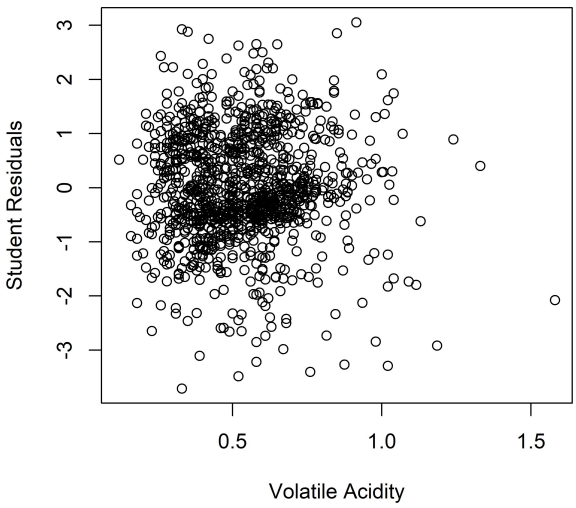

Figure 5: Student residuals vs. X_2.

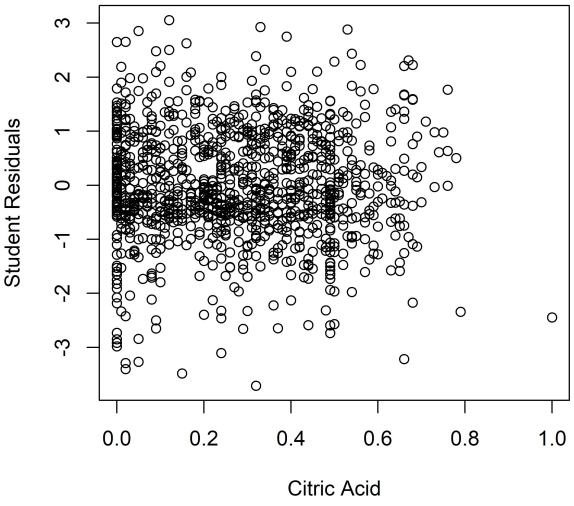

Figure 6: Student residuals vs. X_3.

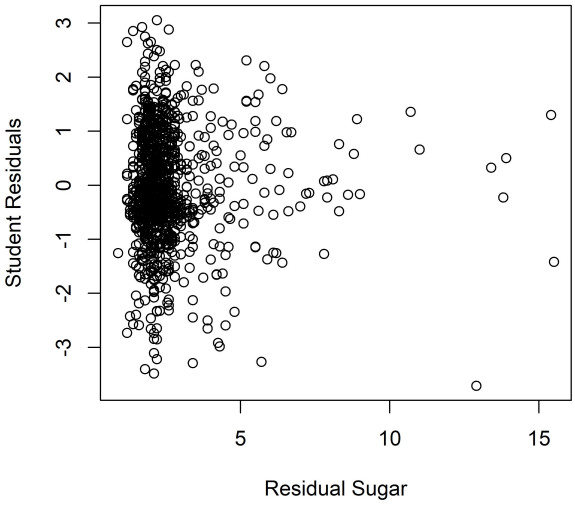

Figure 7: Student residuals vs. X_4.

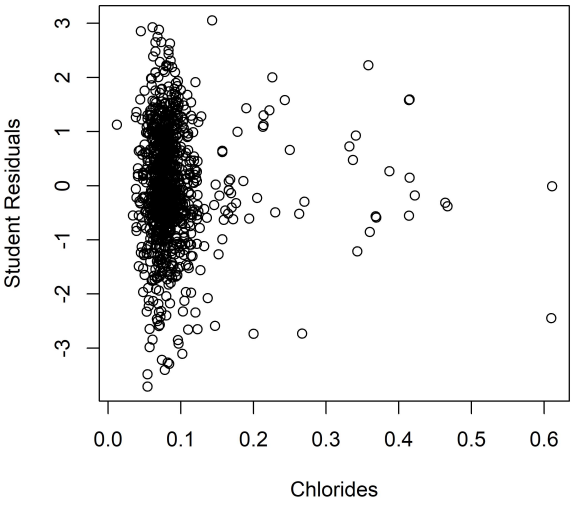

Figure 8: Student residuals vs. X_5.

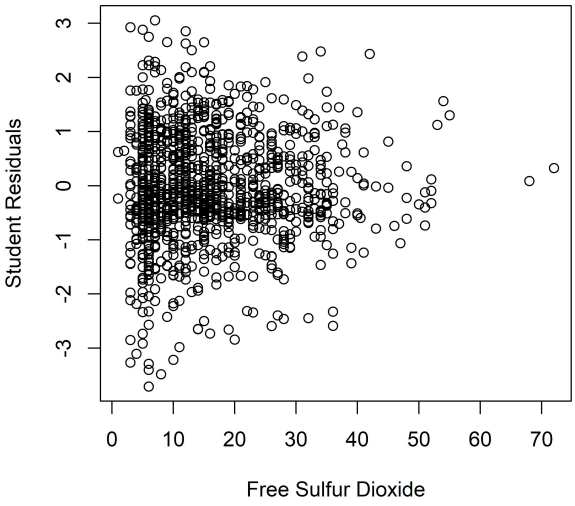

Figure 9: Student residuals vs. X_6.

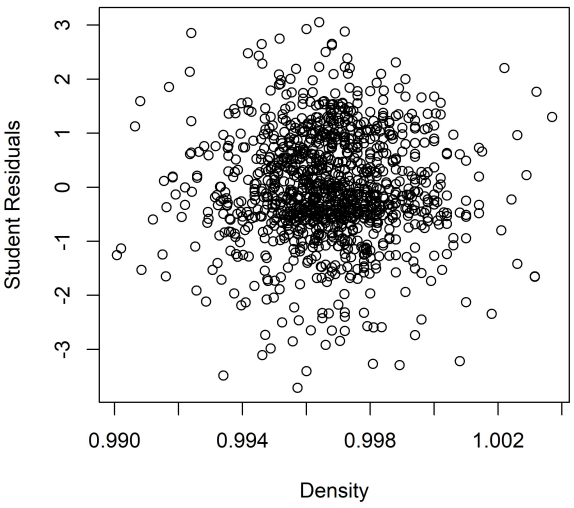

Figure 10: Student residuals vs. X_7.

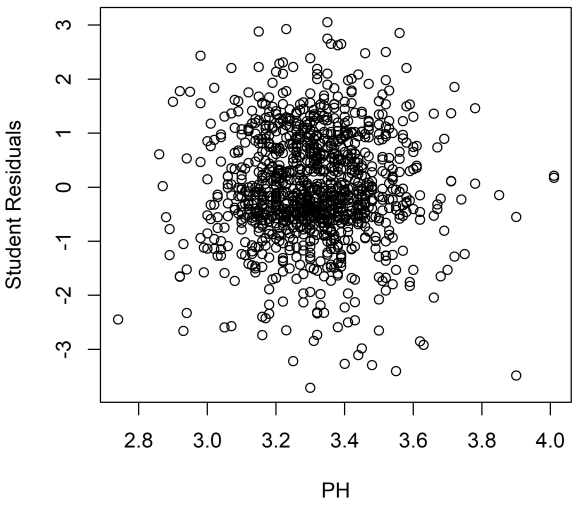

Figure 11: Student residuals vs. X_8.

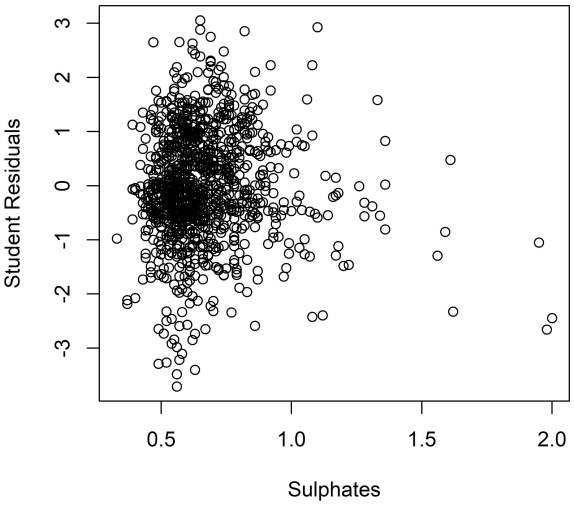

Figure 12: Student residuals vs. X_9.

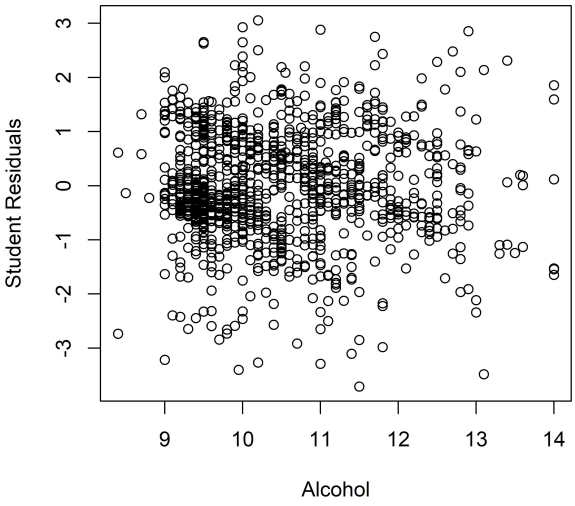

Figure 13: Student residuals vs. X_{10}.

Figure 1 shows that the model's residuals do not violate any of the assumptions for a simple linear regression model. The means of the residuals looks to be very close to 0 and there is not pattern appearing confirming that variance is constant. In Figure 2 the residuals also seem spread out but diagonal lines appear in this graph. We believe these lines appear because our response variable quality is ordinal. The assumption that the variance is normal is being violated as shown by the QQ-plot it appears to be close to the normal distribution but strays away at both tail ends of the data. For the plots of the residuals against all the regressors in the model we noticed figure 4, 5, 6, 10, 11, and 13 for

fixed acidity, volatile acidity, citric acid, density, PH, and alcohol did not look to violate any assumptions although there did appear to be some possible outliers showing in these graphs particularly on the right sides of figure 4 and 5. In figures 7, 8, 9, and 12 there appears to be a fan shape which would violate the assumption of constant variance. It was difficult to tell if there were clusters of outliers or if assumptions were being violated.

Based on what we saw we decided it would be best to try some transformations on the model variables. First we focused on transforming the response variable and used the Box-Cox method to find an optimal transformation although our results from this indicated Y^1 was ideal so we had also tried a log transformation on Y that was also unsatisfactory. Finally we tried log transformations on the variables that had fan shapes in the regressor vs. residual plots and this seemed to improve the model adequacy.

2.2.2 Model Transformation

Simple Linear Regression Model Transformed

$$Y = X_1 + X_2 + X_3 + log(X_4) + log(X_5) + log(X_6) + X_7 + X_8 + log(X_9) + X_{10} + \epsilon$$

Fitted Simple Linear Regression Model Transformed

$$\begin{aligned} Y =& 42.90 + 0.07X_1 - 1.15X_2 - 0.40X_3 + 0.10log(X_4) - 0.21log(X_5) - 0.02log(X_6) \\ & - 39.56X_7 - 0.25X_8 + 0.74log(X_9) + 0.28X_{10} \end{aligned}$$

R^2	R^2_{adj}	MSRES
0.3823	0.3774	0.409

In comparison the transformed model has a lower MSRES and higher R^2 and R^2_{adj} value. Again the model adequacy needs to be checked.

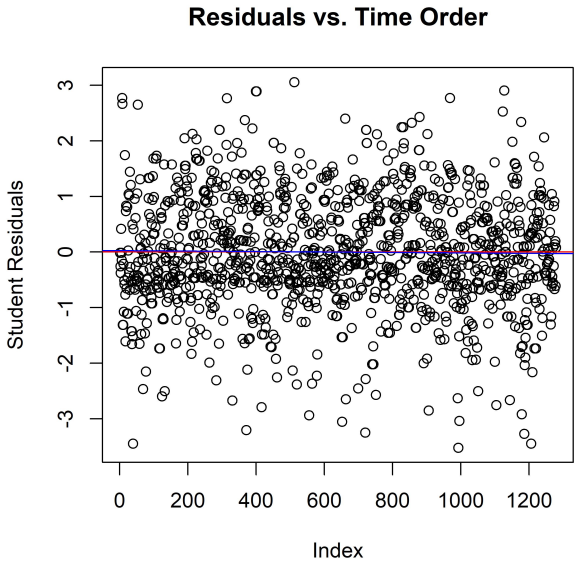

Figure 14: Student residuals plot.

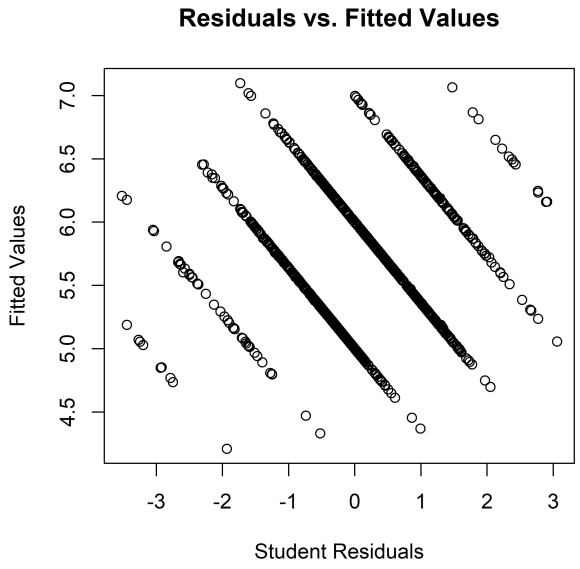

Figure 15: Student residuals vs. fitted values.

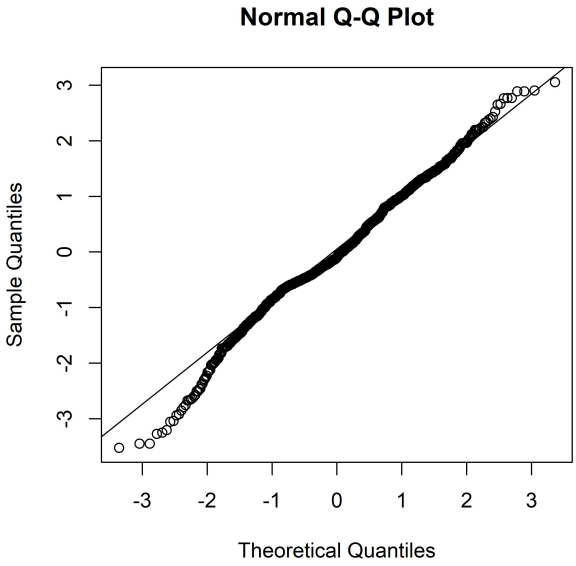

Figure 16: QQ-plot.

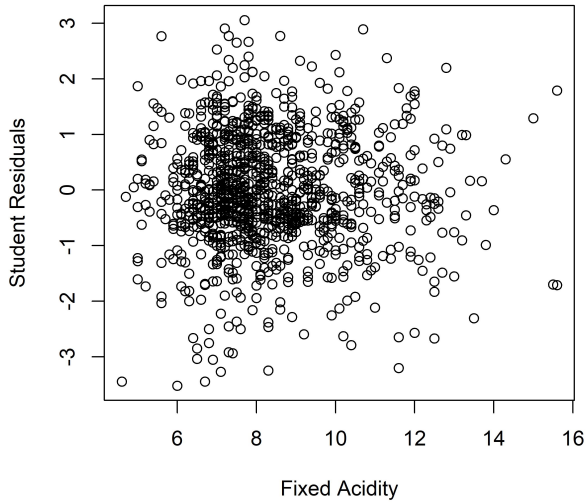

Figure 17: Student residuals vs. X_1.

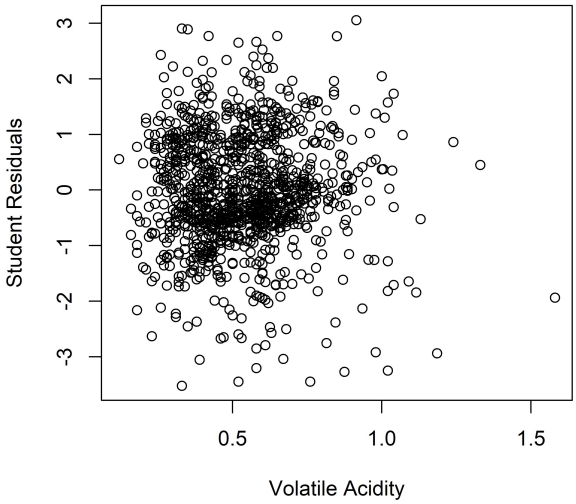

Figure 18: Student residuals vs. X_2.

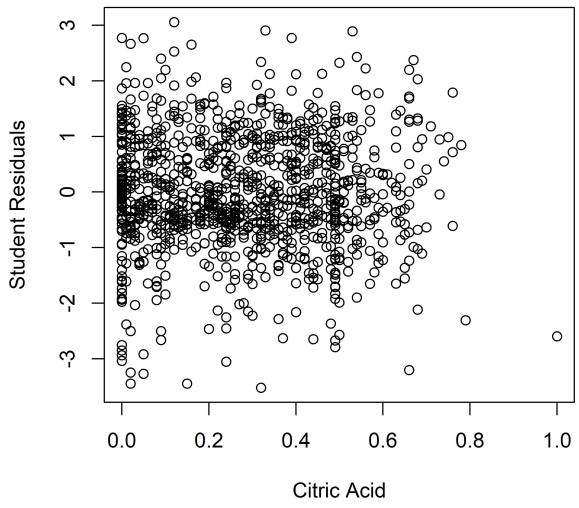

Figure 19: Student residuals vs. X_3.

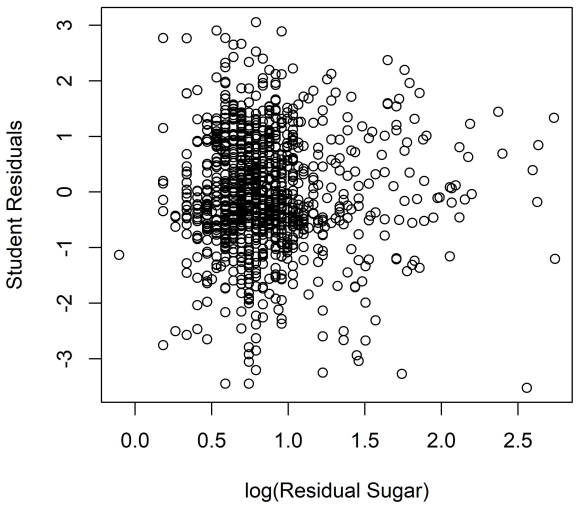

Figure 20: Student residuals vs. X_4.

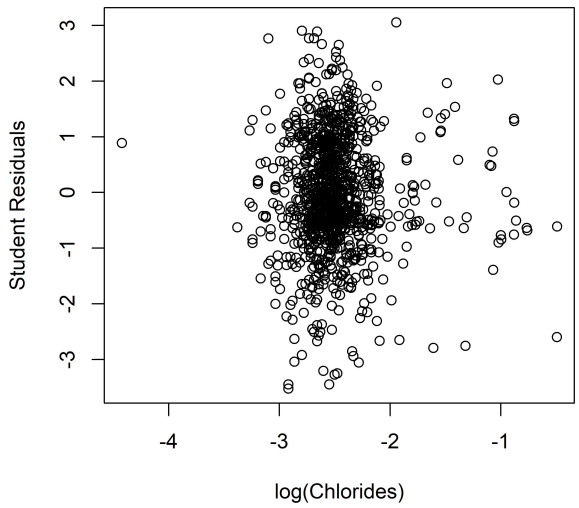

Figure 21: Student residuals vs. X_5.

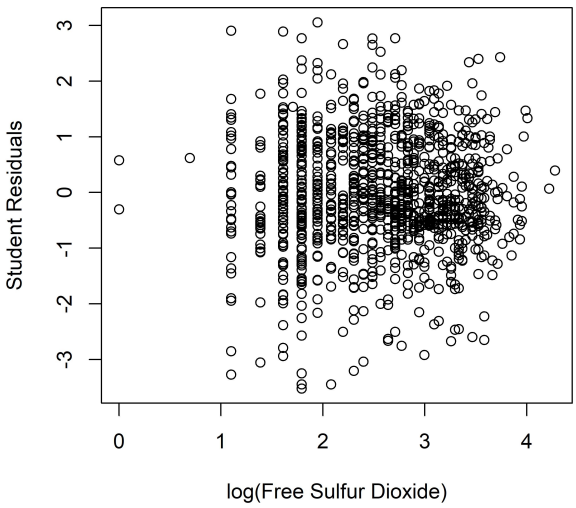

Figure 22: Student residuals vs. X_6.

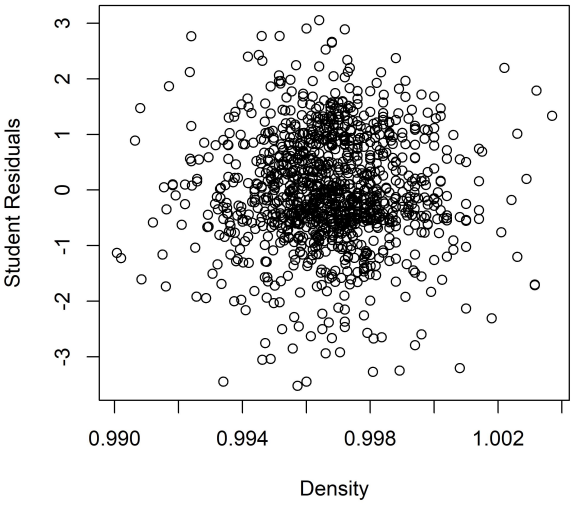

Figure 23: Student residuals vs. X_7.

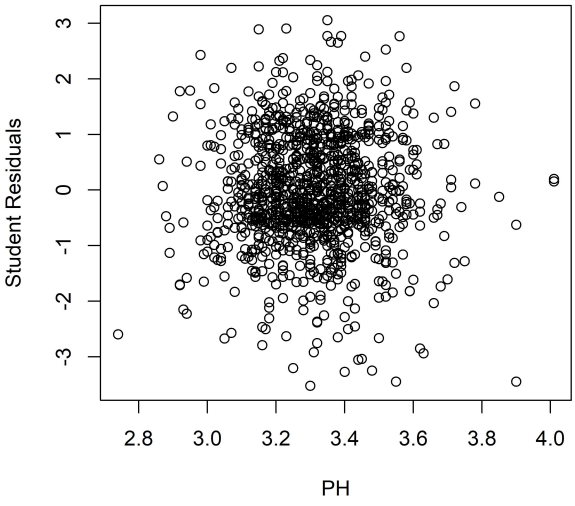

Figure 24: Student residuals vs. X_8.

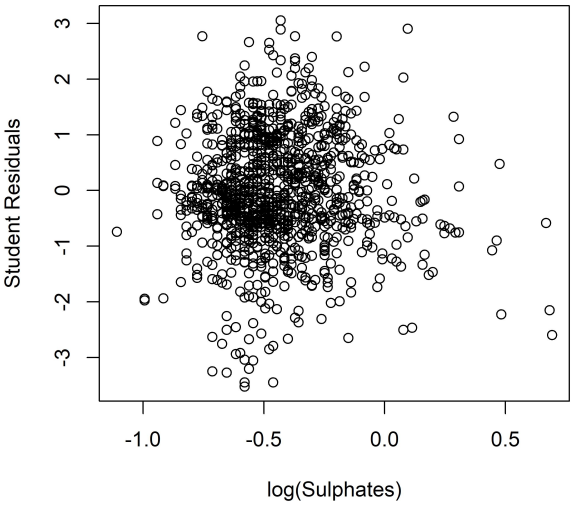

Figure 25: Student residuals vs. X_9.

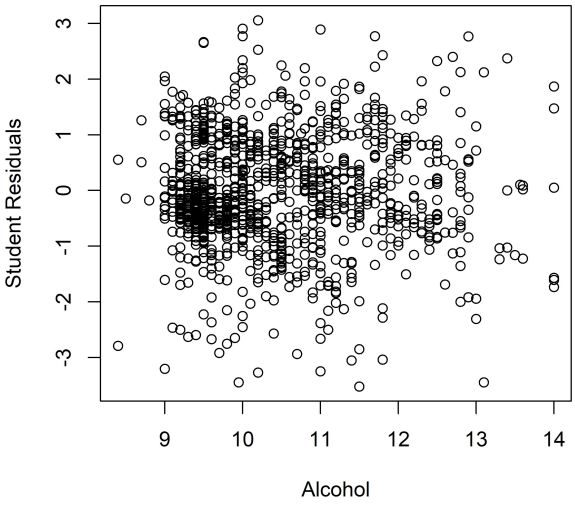

Figure 26: Student residuals vs. X_{10}.

Now in all graphs there are no assumptions being violated except the QQ-plot in figure 16 which has not improved.

2.3 Method A

To find a best fit model that can successfully predict wine quality it was decided a polynomial of order 2 regression model might be better suited. Using the knowledge from

the first model this model was also fitted with the transformed variables. The correlation between all the regressor variables and their squares was found to be quite high so all the variables were centered which reduced the correlation drastically. The maximum correlation between two different variables after centering was 0.6850119 and before centering was 0.9999986 so this was a great improvement. New notation is introduced below because the centered variables are used throughout Method A as well as the ones that have been transformed using log.

$$
\begin{aligned}
X_{1C} &= X_1 - mean(X_1) & X_{2C} &= X_2 - mean(X_2) \\
X_{3C} &= X_3 - mean(X_3) & X_{4CL} &= log(X_4) - mean(log(X_4)) \\
X_{5CL} &= log(X_5) - mean(log(X_5)) & X_{6CL} &= log(X_6) - mean(log(X_6)) \\
X_{7C} &= X_7 - mean(X_7) & X_{8C} &= X_8 - mean(X_8) \\
X_{9CL} &= log(X_9) - mean(log(X_9)) & X_{10C} &= X_{10} - mean(X_{10})
\end{aligned}
$$

Full Polynomial Model A

$$
\begin{aligned}
Y = {} & X_{1C} + X_{1C}^2 + X_{2C} + X_{2C}^2 + X_{3C} + X_{3C}^2 + X_{4CL} + X_{4CL}^2 + X_{5CL} + X_{5CL}^2 \\
& + X_{6CL} + X_{6CL}^2 + X_{7C} + X_{7C}^2 + X_{8C} + X_{8C}^2 + X_{9CL} + X_{9CL}^2 + X_{10C} + X_{10C}^2 + \epsilon
\end{aligned}
$$

Intercept Only Model A

$$Y = 1$$

These two models will be used for variable selection because not all the variables need to be used in the final model. Variable selection was performed with forward, backwards and stepwise methods. Forward selection started with the intercept only model adding in variable terms one at a time based on which had the highest F statistic value. Once the p-value for the F statistic is greater than 0.05 for all remaining regressors not in the model no more regressors will be added into the model. Backwards selection works in a similar way but starting with the full polynomial model and removing terms with the lowest F statistic. The backwards selection method is terminated when the p-value for the F statistic is less than 0.05 for all regressors in the model. Stewise regression method starts with the intercept only model and uses both forwards and backwards methods to add one regressor then check if any need to be removed then add this method terminates once the p-value for the F statistic is greater than 0.05 for all remaining regressors not in the model. In this case all three methods of variable selection returned the same model which will be called Model A.

2.4 Model A

$$Y = X_{2C} + X_{3C} + X_{5CL} + X_{8C} + X_{9CL} + X_{9CL}^2 + X_{10C} + \epsilon$$

Fitted Model A

$$Y = 5.71 - 1.04X_{2C} - 0.33X_{3C} - 0.20X_{5CL} - 0.82X_{8C} + 0.98X_{9CL} - 1.27X_{9CL}^2 + 0.32X_{10C}$$

R^2	R^2_{adj}	MSRES
0.3972	0.3939	0.399

This model has a much higher R^2 and R^2_{adj} values than the previous model and also a much lower MSRES.

2.5 Method B

New notation is introduced below that is used throughout Method B.

$$X_{4log} = log(X_4)$$
$$X_{5log} = log(X_5)$$
$$X_{6log} = log(X_6)$$
$$X_{9log} = log(X_9)$$

Full model

$$\begin{aligned} Y =& 42.9 + 0.07X_1 - 1.15X_2 - 0.40X_3 + 0.096X_{4log} - 0.21X_{5log} \\ & - 0.02X_{6log} - 39.56X_7 - 0.25X_8 + 0.74X_{9log} + 0.28X_{10} \end{aligned}$$

Again, not all the variables need to be used in the final model. The approach using stepwise variable selection is performed the same as Model A. As a result, Alcohol (X_{10}), volatile.acidity (X_2), sulphates (X_9), chlorides(X_5) and pH (X_8) are selected.

Model B with selected regressors

$$Y = 4.15 - 1.003X_2 - 0.25X_{5log} - 0.48X_8 + 0.68X_{9log} + 0.32X_{10}$$

Full Polynomial Model B with selected regressors

A polynomial of order 2 regression model is applied on the selected regressors.

$$\begin{aligned} Y =& 5.66 - 5.52X_2 - 0.97X_2^2 - 2.62X_{5log} + 0.62X_{5log}^2 - 3.63X_8 \\ & - 1.47X_8^2 + 5.65X_{9log} - 3.84X_{9log}^2 + 12.22X_{10} - 0.103X_{10}^2 \end{aligned}$$

2.6 Model B

Reduced Polynomial Model B with selected regressors

Polynomial of order 2 were kept in the model based on which variables were significant in the summary results producing the final Model B.

$$Y = 2.26 - 0.87X_2 - 0.22X_{5log} - 3.57X_8 - 1.40X_8^2 + 5.75X_{9log} - 3.72_{9log}^2 + 0.32X_{10}$$

R^2	R^2_{adj}	MSRES
0.3963	0.393	0.399

2.7 Outliers

The residual graphs indicate there might be potential outliers in the data. To investigate we used the diagonals of the H matrix defined below.

$$H = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

Where $\mathbf{X} = [X_{2C}, X_{3C}, X_{5CL}, X_{8C}, X_{9CL}, X^2_{9CL}, X_{10C}]$ There are potential influence points wherever $2(k+1)/n < \mathbf{H}_{diagonal}$ where k is the number of columns in the $\mathbf{X}$ matrix in this case k=7 and n is the number of rows in the $\mathbf{X}$ matrix in this case n = 1279. For Model A there were 84 influence points found which was too many to try every combination of exclusion/inclusion in the model. One version of Model A, Model A None, was fit with all 84 influence points removed and another version of Model A, Model A Some, was fit with all the influence points removed that when removed by themselves both increased the R^2 value and decreased the MSRES value.

	R^2	R^2_{adj}	MSRES
Model A	0.3972	0.3939	0.399
Model A None	0.3888	0.3852	0.383
Model A Some	0.3673	0.3636	0.396

Removing the outliers decreased the R^2 value but also decreased the MSRES value so it was determined that the outliers should stay in the model. It seems cluster analysis might be useful to identify if the large group of outliers is actually telling us something about the data.

2.8 Model Evaluation

For training set, Ajusted R-squared and MSRES are used to evaluate the models. For test set, new R-squared, Mean Squared Prediction Error (MSPE) and accuracy percentage are used to evaluate the models. The accuracy percentage comes from a confusion matrix used on both Model A and Model B. Because the response variable quality is ordinal, in order to use the confusion matrix the predicted values are rounded up and down so they can be compared with the quality test values.

3 Results & Discussion

We choose 80% of the dataset as training data and 20% as testing data.
As we can see from the experimental results below, both model A and model B gives the same R^2_{adj} and $MSRES$ while the results on test set shows that model A performs slightly better than model B.

	Train Set		Test Set		
Model	R^2_{adj}	$MSRES$	R^2	MSPE	Accuracy
A	0.393	0.399	0.287	0.4478	57.81%
B	0.393	0.399	0.284	0.4495	57.19%

Confusion Matrix Model A

Prediction	Actual					
	3	**4**	**5**	**6**	**7**	**8**
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	9	95	38	1	0
6	1	3	49	77	24	3
7	0	0	1	5	13	0
8	0	0	1	0	0	0

Confusion Matrix Model B

Prediction	Actual					
	3	**4**	**5**	**6**	**7**	**8**
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	9	97	42	2	0
6	1	3	47	74	24	3
7	0	0	1	4	12	0
8	0	0	1	0	0	0

A concern for Model A and Model B is the low R^2 and R^2_{adj} values. The low values indicate that the regressor variables are not able to explain much of the variation in the response variable. This could be due to a need for a non-linear model or a lack of regressor variables that relate to the response variable. Due to the high number of regressor variables we decided not to include any interaction terms which if included could have potentially helped to increase the R^2 and R^2_{adj} values.

Comparing both confusion matrices it appears that both models predict the wine quality in a similar way. It also appears to be relatively accurate as most predicted values are at least close to the actual values even when they are not correct. In conclusion Model A performs slightly better than Model B but with both models having about 57% prediction accuracy either could be used safely to predict wine quality.

Possible directions for future work is to use methods that don't need assumptions such as using Generalized Linear Model or using machine learning algorithms such as K Nearest Neighbors Regression. It would also be recommended to further examine the outliers with cluster analysis because there were such a high number of influence points. Besides, because the response variable, quality, can be assumed to be an ordinal variable and not a continuous variable, the wine quality prediction task can also be used as a multi class classification task.

References

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties. in decision support systems," 2009.