IIE 578: Regression Analysis
Project: Multiple Regression model for Wine Quality Data Set

Vishal Kumar
Arizona State University
1215200480
vkumar47@asu.ed

Introduction:

The project is regarding finding a linear relation between a number of regressors and a response variable. In this project, I studied over the red vinho verde wine samples dataset provided by Paulo Cortez, University of Minho, Guimarães, Portugal. The description of the data is given below. For this project I worked towards building a multiple linear regression model for this dataset. Out of the total number of samples provided I kept 70% for the training-set, 20% for the validation-set and 10% for the test-set. The software in use is JMP[3] pro 14 on windows 10 platform.

Data-Description:

The dataset includes red vinho verde wine samples, from the north of Portugal. It's a multivariate dataset with 1599 data samples.

Data Set Characteristics:	Multivariate
Attribute Characteristics:	Real
Associated Tasks:	Regression
Number of Instances:	1599
Number of Attributes:	12
Missing Values?	None
Area:	Business
Date Donated:	2009-10-07
Number of Web Hits:	1039809

Table 1: Dataset Description

This dataset can be viewed as classification or regression task. The classes are ordered and not balanced (e.g. there are many more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, it has been mentioned that it is not necessary that all input variables are relevant or not.

Attribute Information:
Input variables (based on physicochemical tests):
1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density

9 - pH
10 - sulphates
11 - alcohol
Output variable (based on sensory data):
12 - quality (score between 0 and 10)

A sample from the dataset is shown below:

fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
8	0.52	0.25	2	0.078	19	59	0.99612	3.3	0.48	10.2	5
7	0.23	0.4	1.6	0.063	21	67	0.9952	3.5	0.63	11.1	5
6.9	0.63	0.33	6.7	0.235	66	115	0.99787	3.22	0.56	9.5	5
6.1	0.64	0.02	2.4	0.069	26	46	0.99358	3.47	0.45	11	5
6.5	0.63	0.33	1.8	0.059	16	28	0.99531	3.36	0.64	10.1	6
7.2	0.695	0.13	2	0.076	12	20	0.99546	3.29	0.54	10.1	5
10.8	0.4	0.41	2.2	0.084	7	17	0.9984	3.08	0.67	9.3	6
9.6	0.32	0.47	1.4	0.056	9	24	0.99695	3.22	0.82	10.3	7

Table 2: Sample from red vinho verde wine samples dataset.

Upon careful examination I found that the data contains no missing values nor any inappropriate values like the presence of some string, no preprocessing was required.

Dataset source:
Paulo Cortez, University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez
Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal @2009 [1].

Modelling process:

I chose to work on an order-1 Multiple Linear Regression model and did not include any second-order terms in the process.

I worked on a couple of approaches for this project including building the all-possible-models approach. But here I'm explaining the process that I myself found most intuitive to work with, the intermediate models in this process has been talked about as well.

My approach was to use the backward elimination technique starting with all regressors and only keep the regressors for which the p-value in the t-test is less than the α which is 0.05. This is shown under the 'Parameter Estimates' tab in the JMP.

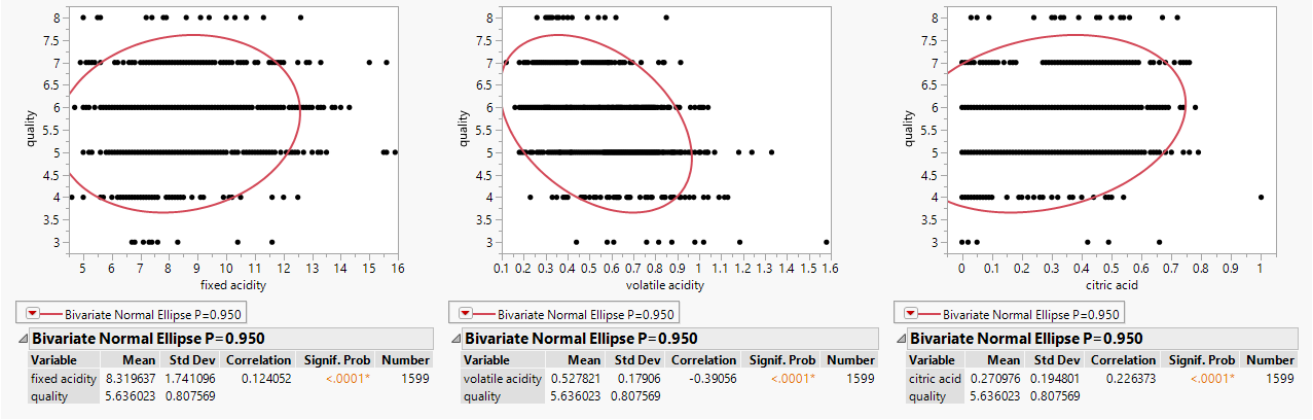
Along with this, I had to make sure that the Root Mean Square Error and R^2_{adj} decreases and increases respectively or at least does not increase/drop significantly in the subsequent iterations. Both values for the model can be seen under 'Summary of Fit' tab.

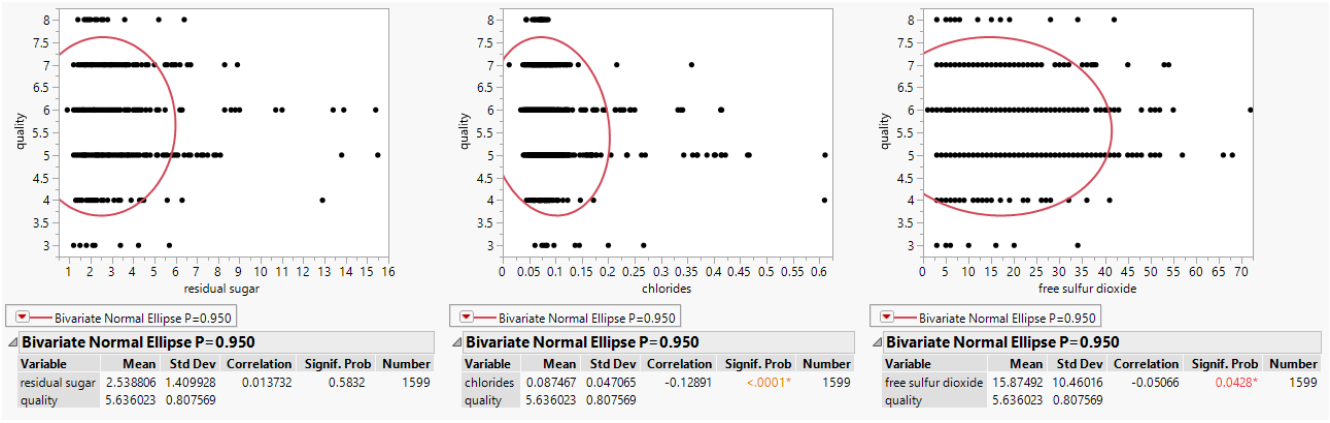
Before finalizing the model, I analyzed the different residual plot and made sure that the F-statistic for Analysis of variance is appropriate.

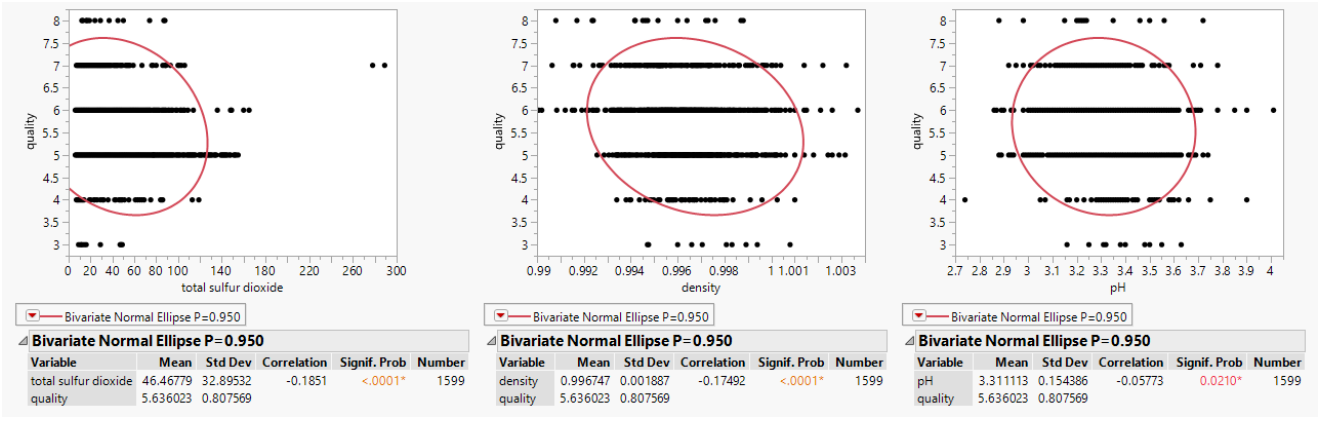
For this project, I did not write the SAS code or any script. I performed all the operations using drag and drop options in JMP pro 14.

I started by examining regressors individually against the response variable to see the individual linear relationship between them. Along with this I drew the density ellipse for them with the density ellipse probability = 0.95.

The results for the above operations are as follow:







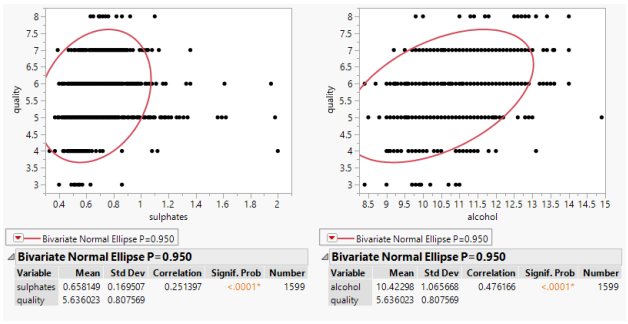


Figure 1: Y vs X for 'quality' & individual response variable.

From the above observations, we can see that many of the regressors have a really low correlation coefficient, and thus showing almost no linear relation ship with the response variable 'quality'. We can also see that 'alcohol' has the highest correlation coefficient, which is kind of intuitive.

Nevertheless, I choose to start with all the regressors for the backward elimination process.

Backward Elimination process:

Iteration 1: For the First iteration, the following observation was obtained.

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%	VIF
Intercept	4.2149101	0.458945	9.18	<.0001*	3.3144447	5.1153755	.
alcohol	0.282686	0.01981	14.27	<.0001*	0.2438189	0.3215532	1.2238734
volatile acidity	-1.17458	0.11847	-9.91	<.0001*	-1.407023	-0.942138	1.2198008
total sulfur dioxide	-0.001903	0.00059	-3.23	0.0013*	-0.00306	-0.000746	1.0439403
sulphates	0.7505421	0.125193	6.00	<.0001*	0.5049099	0.9961743	1.2655903
chlorides	-1.689101	0.495766	-3.41	0.0007*	-2.661811	-0.716392	1.286838
pH	-0.355701	0.135723	-2.62	0.0089*	-0.621994	-0.089408	1.248594

RSquare	0.342158
RSquare Adj	0.33872
Root Mean Square Error	0.648241
Mean of Response	5.615584
Observations (or Sum Wgts)	1155

Table 3: Parameter estimates and summary of fit for initial Iteration.

As we can observe that regressor 'density' has the highest p-value, we will remove it in the next iteration and re-fit the model.

Iteration 2:
R^2_{adj}= .3404
Regressor with highest p-value(0.5373) = 'fixed acidity'.

Iteration 3:
R^2_{adj}= .3406
Regressor with highest p-value(0.2496) = 'residual sugar'.

Iteration 4:

R^2_{adj}= .3404
Regressor with highest p-value(0.2708) = ‘citric acid’.

Iteration 5:
R^2_{adj}= .3403
Regressor with highest p-value(0.0525) = ‘free Sulphur dioxide’.

Iteration 6:
R^2_{adj}= .3387
At this stage all the regressors had a p-value less than 0.05 thus contributing to the model.
The VIF values for all regressors were close to 1, indicating no collinearity between the regressors which was a good thing.

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	4.2149101	0.458945	9.18	<.0001*	.
alcohol	0.282686	0.01981	14.27	<.0001*	1.2238734
volatile acidity	-1.17458	0.11847	-9.91	<.0001*	1.2198008
total sulfur dioxide	-0.001903	0.00059	-3.23	0.0013*	1.0439403
sulphates	0.7505421	0.125193	6.00	<.0001*	1.2655903
chlorides	-1.689101	0.495766	-3.41	0.0007*	1.286838
pH	-0.355701	0.135723	-2.62	0.0089*	1.248594

Table 4: Parameter estimates of remaining regressors.

RSquare	0.342158
RSquare Adj	0.33872
Root Mean Square Error	0.648241
Mean of Response	5.615584
Observations (or Sum Wgts)	1155

Table 5: Summary of fit

Source	RSquare	RASE	Freq
Training Set	0.3422	0.64627	1155
Validation Set	0.3595	0.68521	288
Test Set	0.4343	0.58992	156

Table 6: Cross-validation over training, validation and test set.

After this iteration, I observed that removing any variable from the model was only decreasing the model’s R^2_{adj} drastically.

At this point the Root mean square error is 0.6482, thus I further explored the residual plots.

1. Residuals vs predicted ‘quality’ values.

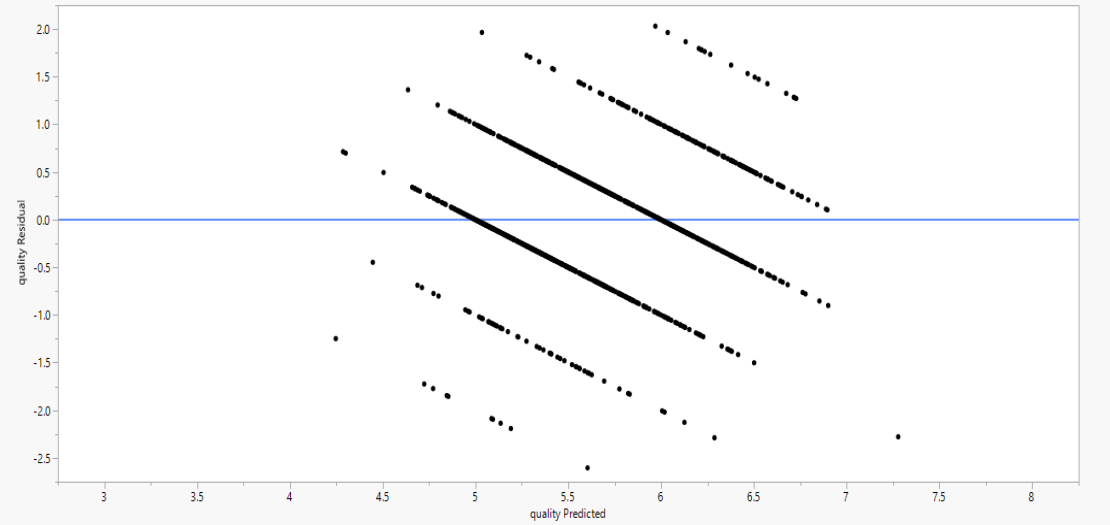


Figure 2: Residuals vs predicted values plot.

At a first glance the plot gives a sense of double bow which sometimes occurs in the case when the Y-values are in the interval of 0 to 1, which is not the case here. I was confused about the final decision of the shape so I referred to the book 'Introduction to linear regression analysis' by Douglas C. Montgomery[2]. I found the similar plot of page 175, figure 5.3, as shown below. Clearly the graph I got may had a similar outline but the internal structure of the graph is not random and thus suggests some transformation over the response variable 'quality'.

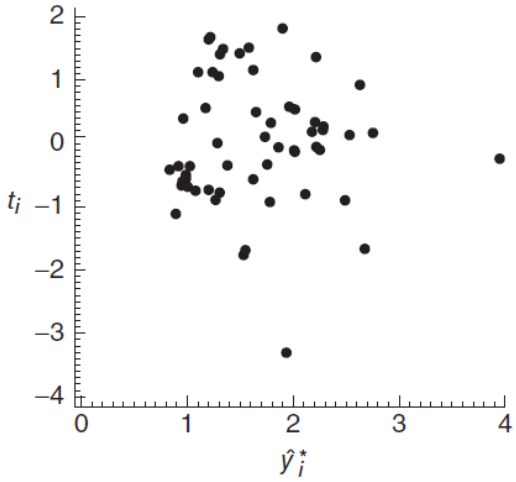


Figure 3: Figure 5.3, page 175, Introduction to Linear Regression Analysis

2. Normal Probability plot:
   The Normal Probability plots gave good results, the distribution of the residuals turned out to be quite 'Normal' in nature, the obtained graphs are as following:

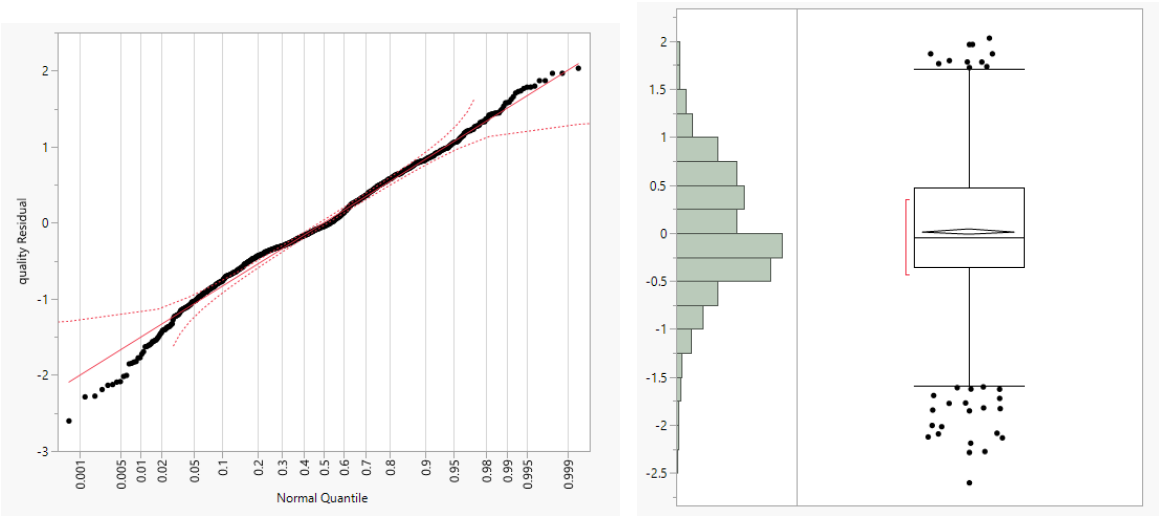


Figure 4: Residual Normal Quantile Plot(Left) & Normal Distribution curve for residuals(Right).

Summary Statistics for the residuals is as following:

Mean	0.0146008
Std Dev	0.6482741
Std Err Mean	0.0162119
Upper 95% Mean	0.0463997
Lower 95% Mean	-0.017198

Table 7: Residuals Summary Statistics

Experiments with Transformation over the response variable 'quality':
I performed the following transformation on 'quality':

1. Arrhenius Inverse
2. Square Root
3. Log(y)
4. Log(y) + 1
5. Square
6. Cube root
7. Cube
8. Exponential
9. Logistic
10. Logistic percent
11. Reciprocal

Unfortunately, none of the above transformation worked in a desire manner, while most gave the similar results as that of without any transformation, some actually gave worse results, leading to the conclusion that the above mentioned transformation are not good for this data.

Arcsine, transformation did not work as values were not between 0 – 1.

Further Experimentation:
I tried a couple of cross-products between some regressors that were already in my model, apparently the combination made were random, and I did not get any significant results using them in my model. Thus, my 2nd-order models did not work.

Result: The best 1-order multiple regression model that I was able to achieve is given by the following equation.

$$quality = 4.21 + 0.28 * alcohol - 1.17 * volatile\ acidity - 0.0019 * total\ sulphur\ dioxide + 0.75 * sulphates - 1.69 * chlorides - 0.35 * PH$$

The Root Mean Square Error while fitting on 1155 samples was 0.64.
And the analysis of variance is shown below:

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	250.91110	41.8185	99.5166
Error	1148	482.40839	0.4202	**Prob > F**
C. Total	1154	733.31948		<.0001*

Table 8: Analysis of variance for the model.

Cross-validation for this model is shown in Table 6.
Out of all the model I tried, this one has the best Press value, details are as follow:

Press	Press RMSE	Press RSquare
682.10157209	0.76848171	0.0698

Table 9: Press Value for the model.

Future Work:
For this dataset, non-linear model may fit better, and one can try the different combinations of significant regressors to increase the effectiveness of the linear model.

Acknowledgement:
I would like to thank Prof. Douglas Montgomery for the guidance, encouragement and motivation for the subject. And I would like to thank my chair of the graduate program in computer science, prof. Chitta Baral to allow me to study a subject out of my expertise.

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, Elsevier, 47(4):547-553, 2009.
[2] Montgomery, Douglas. Introduction to linear regression analysis. Hoboken, John Wiley & Sons, Inc., 2012.
[3] JMP [SAS Institute]. 2014. Retrieved from https://www.jmp.com/en_us/home.html