

# IMSE 586 PROJECT ASSIGNMENT

## Part B: Wine Data Quality Prediction

**Team:**

*Suresh Ooty Krishnaswamy*

*Selvamani Masilamani*

# Contents

Part B: Wine Data Quality Prediction .....	1
Team Effort: .....	3
Data: .....	3
B1: .....	5
B2 .....	11
B3: .....	18
B4: .....	24
B5: .....	28

## TABLES

Table 1 : Comparison of A3 & A4 .....	3
Table 1 : Six Stages .....	16
Table 3 : Six Stages .....	22
Table 3 : Six Stages .....	26
Table 3 : Six Stages .....	32

## FIGURES

Figure 3: A1 regression lines on Scatterplot .....	3
--	---

**FIGURE 1: A1 regression lines on Scatterplot**

**TABLE 1 : COMPARISON OF A3 & A4**

### Team Effort:

As the team comprises of two members, the effort has been equally divided between the members, each contributing to this part of project (Part B) equally at 50%.

**“The Project Team has not given nor received any aid on this assignment”**

### Data:

A mixture of Red and White wine data (a total of 2000 observations) has been made available with one dependent variable (Quality) and 11 independent variables as given below

- 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)

The objective of the study is to identify the variables that affect the quality of wine and if Red and White wine quality is affected by the same predictors at the same degree. In the data set provided, assume that the first 1000 observations are for Red Wine, and the last 1000 observations are for White wine

A sample of data is as follows.

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	Y	Wine Type
1	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5	Red
2	7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5	Red
3	7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5	Red
4	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6	Red
5	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5	Red

6	7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5	Red
7	7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3.3	0.46	9.4	5	Red

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
X2	0.203									
	0									
X3	0.433	-0.419								
	0	0								
X4	-0.199	-0.292	0.175							
	0	0	0							
X5	0.28	0.329	0.076	-0.213						
	0	0	0.001	0						
X6	-0.386	-0.412	0.12	0.535	-0.247					
	0	0	0	0	0					
X7	-0.482	-0.456	0.114	0.579	-0.33	0.803				
	0	0	0	0	0	0				
X8	0.588	0.344	0.159	0.337	0.351	-0.121	-0.213			
	0	0	0	0	0	0	0			
X9	-0.321	0.271	-0.464	-0.29	-0.037	-0.183	-0.269	-0.083		
	0	0	0	0	0.103	0	0	0		
X10	0.368	0.192	0.123	-0.245	0.472	-0.268	-0.331	0.327	0.064	
	0	0	0	0	0	0	0	0	0.004	
X11	0.041	-0.067	0.066	-0.322	-0.164	-0.204	-0.244	-0.493	0.21	0.069
	0.066	0.003	0.003	0	0	0	0	0	0	0.002

From the variable correlation study (method – Pearson), the following sets were identified as correlated.

1. X1 & X8
2. X6 & X7

## B1:

Perform Cluster Analysis on the data given. In this part, assume that you do not know that the first 1000 observations are for Red wine and the last 1000 observations are for White wine. Your Cluster Analysis should include at least the following elements:

- The number of clusters to be tested should include 2, 3, 5 and 11 clusters as a minimum
- Did you try balancing the correlated variables in the set of independent variables used for clustering? Did it help with the explanation of the clusters?
- Did you try standardizing the variables and did it help in explanation of the clusters?
- Can you tell which clusters are for Red and White wines, which clusters are Very Low, Low, Average, Good, and Very Good quality wine?

Show your work and comment on your findings and insights from the results of this step.

Solution:

An approach of taking balanced sample was adopted, as clustering on the 2000 observations does not seem a feasible approach. Two random samples for each values of response were taken for both red and white wine samples. The sample data is as follow. First set comes from first 1000 records ( 6 y values x 2 = 12 records) and the second set comes from second 1000 records ( 7 y values x 2 = 14 records).

Obs	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
460	11.6	0.58	0.66	2.2	0.074	10	47	1.000 8	3.25	0.57	9
518	10.4	0.61	0.49	2.1	0.2	5	16	0.999 4	3.16	0.63	8.4
200	6.9	1.09	0.06	2.1	0.061	12	31	0.994 8	3.51	0.43	11.4
225	8.4	0.635	0.36	2	0.089	15	55	0.997 45	3.31	0.57	10.4
18	8.1	0.56	0.28	1.7	0.368	16	56	0.996 8	3.11	1.28	9.3
22	7.6	0.39	0.31	2.3	0.082	23	71	0.998 2	3.52	0.65	9.7
134	6.6	0.5	0.01	1.5	0.06	17	26	0.995 2	3.4	0.58	9.8
135	7.9	1.04	0.05	2.2	0.084	13	29	0.995 9	3.22	0.55	9.9
359	11.9	0.43	0.66	3.1	0.109	10	23	1	3.15	0.85	10.4
365	12.8	0.615	0.66	5.8	0.083	7	42	1.002 2	3.07	0.73	10
482	9.4	0.3	0.56	2.8	0.08	6	17	0.996 4	3.15	0.92	11.7
496	10.7	0.35	0.53	2.6	0.07	5	16	0.997 2	3.15	0.65	11
1294	9.1	0.59	0.38	1.6	0.066	34	182	0.996 8	3.23	0.38	8.5
1445	7.1	0.32	0.32	11	0.038	16	66	0.993 7	3.24	0.4	11.5
1189	6.5	0.28	0.28	8.5	0.047	54	210	0.996 2	3.09	0.54	8.9
1204	5.8	0.28	0.35	2.3	0.053	36	114	0.992 4	3.28	0.5	10.2
1038	7.3	0.24	0.39	17.95	0.057	45	149	0.999 9	3.21	0.36	8.6

1039	7.3	0.24	0.39	17.95	0.057	45	149	0.999 9	3.21	0.36	8.6
1016	6.3	0.48	0.04	1.1	0.046	30	99	0.992 8	3.24	0.36	9.6
1018	7.4	0.34	0.42	1.1	0.033	17	171	0.991 7	3.12	0.53	11.3
1076	7.1	0.18	0.36	1.4	0.043	31	87	0.989 8	3.26	0.37	12.7
1077	7	0.32	0.34	1.3	0.042	20	69	0.991 2	3.31	0.65	12
1442	6	0.25	0.28	2.2	0.026	54	126	0.989 8	3.43	0.65	12.9
1598	5.9	0.27	0.29	11.4	0.036	31	115	0.994 9	3.35	0.48	10.5
1827	7.4	0.24	0.36	2	0.031	27	139	0.990 55	3.28	0.48	12.5
1876	6.9	0.36	0.34	4.2	0.018	57	119	0.989 8	3.28	0.36	12.7

“Ward Linkage method with Euclidean distance” trials were attempted with Standardized & Non-Standardized values, with a combination of with correlated variables and without correlated variables.

1 WARD linkage, Euclidean distance with all variables: [2 cluster]

Final Partition  
Number of clusters: 2

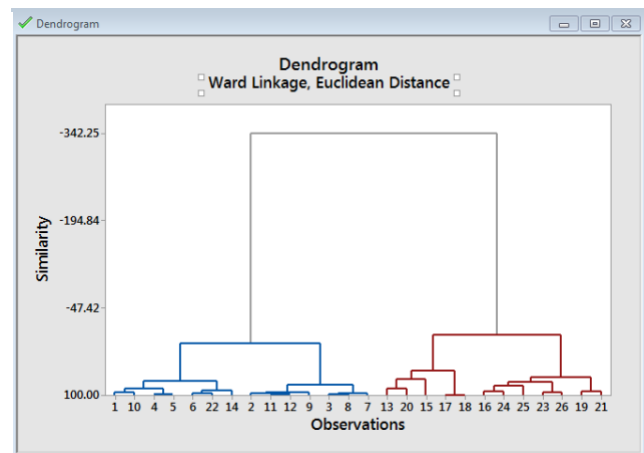
	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	14	5995.4	19.0467	32.5099
Cluster2	12	16442.5	32.9129	73.4375

Cluster Centroids

Variable	Cluster1	Cluster2	Grand centroid
X1	9.0286	6.917	8.0538
X2	0.5529	0.312	0.4419
X3	0.3779	0.323	0.3527
X4	3.0500	5.975	4.4000
X5	0.1029	0.043	0.0751
X6	12.5000	38.417	24.4615
X7	40.2857	138.333	85.5385
X8	0.9971	0.994	0.9955
X9	3.2536	3.248	3.2512
X10	0.6757	0.448	0.5704
X11	10.3214	10.583	10.4423
Y	5.4286	6.167	5.7692

Distances Between Cluster Centroids

	Cluster1	Cluster2
Cluster1	0.000	101.483
Cluster2	101.483	0.000



NOTES:

1. The clusters were not clear.
2. Further, one more cluster tried by dropping variable X7 and standardizing the variables.

## 2. Ward Linkage – Euclidean , with Standardized variables and dropping correlated variable X7

Final Partition  
Number of clusters: 2

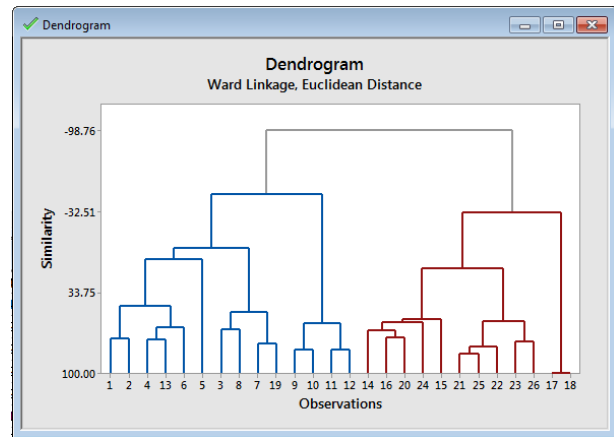
	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	14	134.842	2.98286	5.04810
Cluster2	12	78.805	2.49101	3.56888

### Cluster Centroids

Variable	Cluster1	Cluster2	Grand centroid
X1	0.548280	-0.639660	-0.0000000
X2	0.616223	-0.718927	0.0000000
X3	0.044835	-0.052308	0.0000000
X4	-0.417128	0.486650	-0.0000000
X5	0.432182	-0.504212	0.0000000
X6	-0.619293	0.722508	0.0000000
X8	0.512186	-0.597550	0.0000000
X9	-0.027757	0.032383	0.0000000
X10	0.399329	-0.465883	0.0000000
X11	-0.357205	0.416739	0.0000000
Y	-0.214101	0.249784	0.0000000

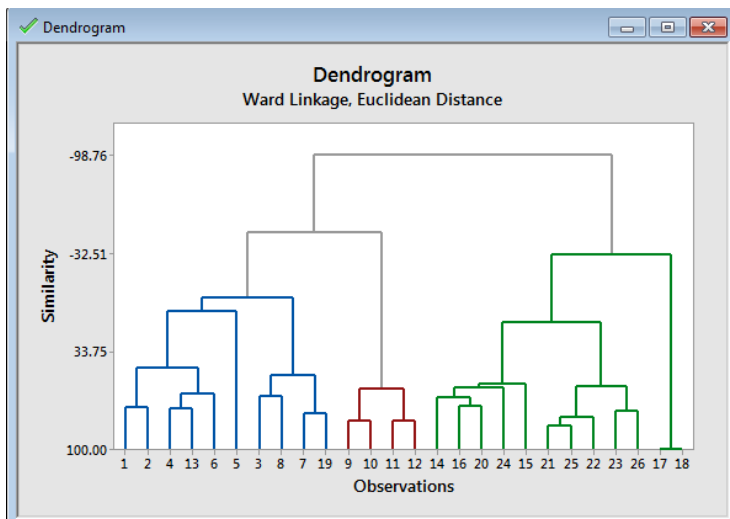
### Distances Between Cluster Centroids

	Cluster1	Cluster2
Cluster1	0.00000	3.08143
Cluster2	3.08143	0.00000



Notes : using this approach , the RED and WHITE wine observations separated neatly.

## 3. WARD Linkage, Standardized variables, dropped X7 [3 cluster]



### NOTES:

1. RED wine observations separated in to LOW quality and HIGH Quality, with some error / noise in the cluster (observations 13, 19)
2. Observations 9, 10, 11, 12 have Y values 7 & 8. Good Red wines
3. All other white wines are in the third cluster.

Final Partition  
Number of clusters: 3

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	10	88.0993	2.78187	5.18040
Cluster2	4	7.7341	1.33903	1.77472
Cluster3	12	78.8045	2.49101	3.56888

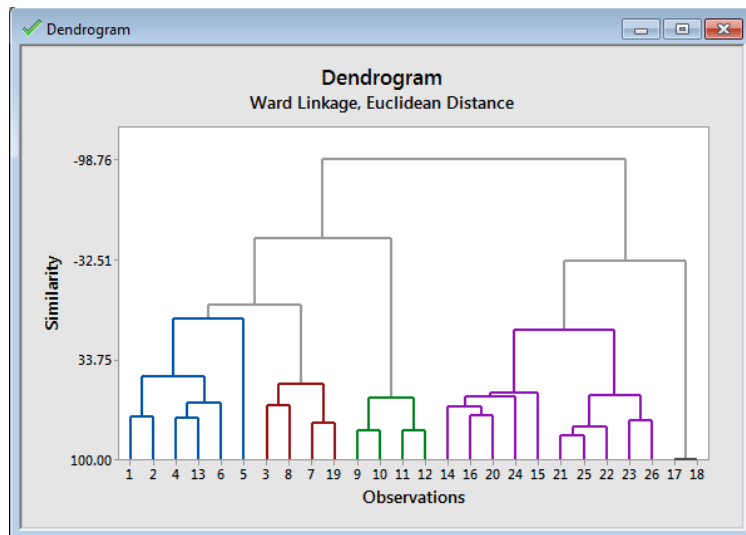
#### Cluster Centroids

Variable	Cluster1	Cluster2	Cluster3	Grand centroid
X1	0.121282	1.61578	-0.639660	-0.0000000
X2	0.894336	-0.07906	-0.718927	0.0000000
X3	-0.495709	1.39620	-0.052308	0.0000000
X4	-0.516361	-0.16905	0.486650	-0.0000000
X5	0.545269	0.14946	-0.504212	0.0000000
X6	-0.432788	-1.08556	0.722508	0.0000000
X8	0.347247	0.92453	-0.597550	0.0000000
X9	0.369165	-1.02006	0.032383	0.0000000
X10	0.142165	1.04224	-0.465883	0.0000000
X11	-0.593920	0.23458	0.416739	0.0000000
Y	-0.659430	0.89922	0.249784	0.0000000

#### Distances Between Cluster Centroids

	Cluster1	Cluster2	Cluster3
Cluster1	0.00000	3.69498	3.16927
Cluster2	3.69498	0.00000	4.22784
Cluster3	3.16927	4.22784	0.00000

4. WARD Linkage, Standardized variables, dropped X7 [5 cluster]:



#### NOTES:

1. Observations 17 , 18 were the mid quality WHITE wines with Y = 5.
2. WHITE wine cluster subdivided in to 2 clusters
3. RED wine clusters are in to three groups ( quality LOW, MEDIUM and HIGH)



Number of clusters: 5

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	6	45.6605	2.56464	4.64024
Cluster2	4	14.6746	1.88941	2.32543
Cluster3	4	7.7341	1.33903	1.77472
Cluster4	10	48.2359	2.12765	3.29556
Cluster5	2	0.0000	0.00000	0.00000

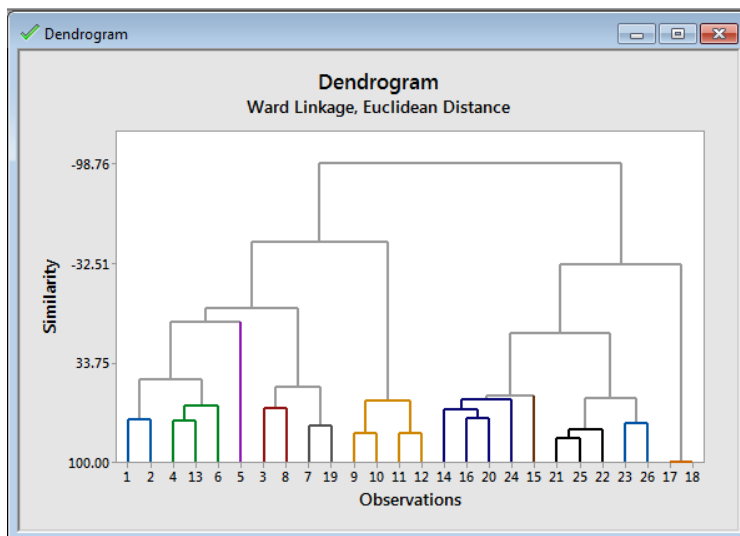
#### Cluster Centroids

Variable	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Grand centroid
X1	0.58863	-0.57974	1.61578	-0.690161	-0.38715	-0.0000000
X2	0.51730	1.45988	-0.07906	-0.687024	-0.87844	0.0000000
X3	0.33893	-1.74766	1.39620	-0.104473	0.20852	0.0000000
X4	-0.49519	-0.54812	-0.16905	0.028687	2.77646	-0.0000000
X5	1.02743	-0.17797	0.14946	-0.552908	-0.26073	0.0000000
X6	-0.45351	-0.40170	-1.08556	0.611641	1.27684	0.0000000
X8	0.73301	-0.23139	0.92453	-0.953342	1.18141	0.0000000
X9	0.10255	0.76909	-1.02006	0.108159	-0.34650	0.0000000
X10	0.52620	-0.43388	1.04224	-0.357075	-1.00993	0.0000000
X11	-0.86421	-0.18848	0.23458	0.759893	-1.29903	0.0000000
Y	-1.00580	-0.13988	0.89922	0.379672	-0.39965	0.0000000

#### Distances Between Cluster Centroids

	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Cluster1	0.00000	3.40124	3.19450	3.92046	4.64686
Cluster2	3.40124	0.00000	5.07602	3.30982	5.29924
Cluster3	3.19450	5.07602	0.00000	4.31916	5.40894
Cluster4	3.92046	3.30982	4.31916	0.00000	4.28266
Cluster5	4.64686	5.29924	5.40894	4.28266	0.00000

#### 4. WARD Linkage, Standardized variables, dropped X7 [11 clusters]:



#### NOTES:

1. When the number of clusters are further added, it could be noted that the groups are formed in terms of similarity of one of the variables.

For example, observation 3 & 8 have similar X11 (alcohol) levels.

Answers to Questions:

Questions	Findings
<ul style="list-style-type: none"> <li>• <b>Did you try balancing the correlated variables in the set of independent variables used for clustering? Did it help with the explanation of the clusters?</b></li> </ul>	<ul style="list-style-type: none"> <li>• When the correlated variables are dropped, the clusters are arranged in meaningful order.</li> <li>• Nevertheless, a combination of dropping correlated variables and standardized did create good set of clusters ( RED &amp; WHITE) wine.</li> </ul>
<ul style="list-style-type: none"> <li>• <b>Did you try standardizing the variables and did it help in explanation of the clusters?</b></li> </ul>	<ul style="list-style-type: none"> <li>• Yes, it helps ordering the clusters and adjusting the similarity among, as the standardization approach shrinks every variable to similar scale.</li> </ul>
<ul style="list-style-type: none"> <li>• <b>Can you tell which clusters are for Red and White wines, which clusters are Very Low, Low, Average, Good, and Very Good quality wine?</b></li> </ul>	<ul style="list-style-type: none"> <li>• Yes, With the approach of 26 samples in a random picked order, the Ward linkage with standardized and independent variables create RED &amp; WINE clusters (2 cluster)</li> <li>• The clusters were exactly not in the order of Very Low, Low, Average, Good and Very Good for all wine types.</li> <li>• A 5 cluster model, did create sub groups , that is more or less of this division.</li> </ul>

## B2:

Perform Regression Analysis using observations 1-500 and 1001 – 1500 data (1000 combined mixed model development data) to find the best predictor model for the quality of wine.

- Comment and show what you have done on each of the six stages of regression analysis and why.
- Compute the accuracy of the regression model by applying the model to the data points 501 – 1000 and 1501 – 2001 (1000 combined test data) and compute the SSE Sum of Squares of Residual (Error) of the test data with the SSE with the model development data.

## Solution:

The train dataset was constructed with 1-500 and 1001-500 and the test dataset was with 501-1000 and 1501-1999.

**Model 1:** Some correlation has been noted between X1 & X8 and X6 & X7. To start with the variable X7 was dropped due to high correlation with X6. The generated model was as follows.

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	10	209.879	20.9879	40.95	0.000
X1	1	6.113	6.1129	11.93	0.001
X2	1	20.236	20.2359	39.48	0.000
X3	1	0.549	0.5493	1.07	0.301
X4	1	0.895	0.8954	1.75	0.187
X5	1	2.173	2.1729	4.24	0.040
X6	1	0.731	0.7314	1.43	0.233
X8	1	2.859	2.8592	5.58	0.018
X9	1	3.860	3.8603	7.53	0.006
X10	1	8.407	8.4070	16.40	0.000
X11	1	23.904	23.9043	46.64	0.000
Error	989	506.872	0.5125		
Lack-of-Fit	845	506.872	0.5998	*	*
Pure Error	144	0.000	0.0000		
Total	999	716.751			

Model Summary			
S	R-sq	R-sq(adj)	R-sq(pred)
0.715898	29.28%	28.57%	27.41%

## Notes:

1. The variables X6, X3, X4 came insignificant
2. The R-square value was 29.28

A final regression equation was achieved by “backward elimination method” manually and using “Stepwise-backward elimination method”.

# Regression Analysis: Y versus X1, X2, X3, X4, X5, X6, X9, X10, X11

## Backward Elimination of Terms

Candidate terms: X1, X2, X3, X4, X5, X6, X9, X10, X11

	-----Step 1-----		-----Step 2-----		-----Step 3-----		-----Step 4-----	
	Coef	P	Coef	P	Coef	P	Coef	P
Constant	1.197		1.034		1.203		1.080	
X1	0.0502	0.011	0.0517	0.009	0.0469	0.009	0.0389	0.014
X2	-1.336	0.000	-1.338	0.000	-1.357	0.000	-1.283	0.000
X3	-0.166	0.383	-0.188	0.319	-0.172	0.356		
X4	-0.00535	0.420						
X5	-1.446	0.009	-1.369	0.013	-1.394	0.011	-1.469	0.007
X6	0.00146	0.393	0.00098	0.540				
X9	0.348	0.058	0.374	0.038	0.352	0.047	0.394	0.021
X10	0.470	0.001	0.475	0.001	0.470	0.001	0.471	0.001
X11	0.3356	0.000	0.3411	0.000	0.3384	0.000	0.3352	0.000
S	0.701836		0.701712		0.701491		0.701439	
R-sq	30.72%		30.67%		30.65%		30.59%	
R-sq(adj)	30.09%		30.11%		30.16%		30.17%	
R-sq(pred)	29.08%		29.18%		29.32%		29.42%	
Mallovs' Cp	10.00		8.65		7.03		5.88	

a to remove = 0.1

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	6	211.652	35.275	70.29	0.000
X1	1	3.042	3.042	6.06	0.014
X2	1	45.565	45.565	90.79	0.000
X5	1	3.704	3.704	7.38	0.007
X9	1	2.741	2.741	5.46	0.020
X10	1	5.567	5.567	11.09	0.001
X11	1	105.227	105.227	209.67	0.000
Error	991	497.355	0.502		
Lack-of-Fit	846	497.355	0.588	*	*
Pure Error	145	0.000	0.000		
Total	997	709.007			

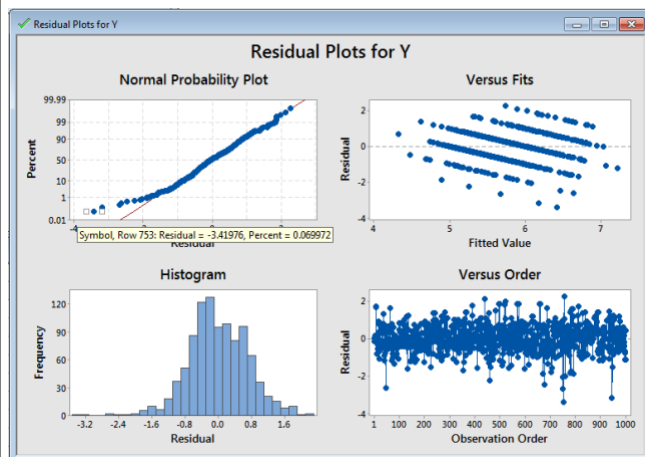
## Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.708429	29.85%	29.43%	28.68%

## Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.082	0.578	1.87	0.061	
X1	0.0392	0.0159	2.46	0.014	1.47
X2	-1.282	0.135	-9.53	0.000	1.29
X5	-1.480	0.545	-2.72	0.007	1.54
X9	0.403	0.172	2.34	0.020	1.44
X10	-0.487	0.146	-3.33	0.001	1.66
X11	0.3306	0.0228	14.48	0.000	1.13

$$Y = 1.051 + 0.0456 X1 - 1.216 X2 - 1.508 X5 + 0.00137 X6 + 0.382 X9 + 0.513 X10 + 0.3273 X11$$



Two outliers were noted (row nos. 753 and 945) and they were dropped. With the new set of data, the same set of variables (from step1) were tried with a "Stepwise-backward elimination method". The set of variables were same, but there was a difference in the coefficients (which could be noted from the regression equation) and in the R2 square value from 29.52 to 30.59.

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	6	214.633	35.772	72.71	0.000
X1	1	2.986	2.986	6.07	0.014
X2	1	45.617	45.617	92.71	0.000
X5	1	3.651	3.651	7.42	0.007
X9	1	2.620	2.620	5.32	0.021
X10	1	5.186	5.186	10.54	0.001
X11	1	107.953	107.953	219.41	0.000
Error	990	487.096	0.492		
Lack-of-Fit	845	487.096	0.576		*
Pure Error	145	0.000	0.000		
Total	996	701.729			

#### Model Summary

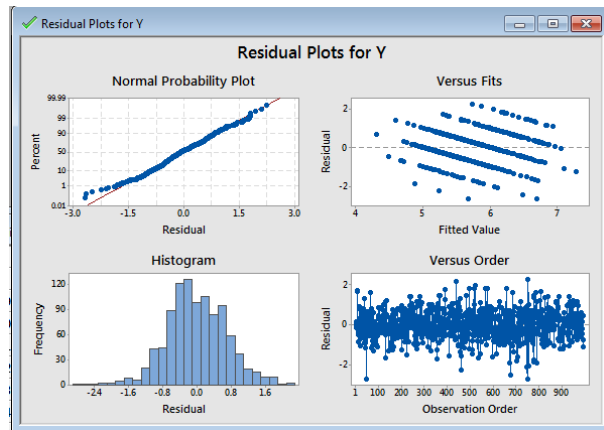
S	R-sq	R-sq(adj)	R-sq(pred)
0.701439	30.59%	30.17%	29.42%

#### Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.080	0.572	1.89	0.059	
X1	0.0389	0.0158	2.46	0.014	1.47
X2	-1.283	0.133	-9.63	0.000	1.29
X5	-1.469	0.539	-2.72	0.007	1.54
X9	0.394	0.171	2.31	0.021	1.44
X10	0.471	0.145	3.25	0.001	1.65
X11	0.3352	0.0226	14.81	0.000	1.13

#### Regression Values:

$$Y = 1.080 + 0.0389 X1 - 1.283 X2 - 1.469 X5 + 0.394 X9 + 0.471 X10 + 0.3352 X11$$



Model 1 SSE : **487.096**

Model 1 prediction on Test dataset: SSE : **587.6833**

All the VIF values are within 10, indicating that there is no multi-collinearity within this model.

**Model 1\_A:** A second model was tried with the Quadratic equation on the independent variables, with variable interactions as well.

The model output was as follows

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	11	236.352	21.4866	45.48	0.000
X1	1	2.746	2.7457	5.81	0.016
X2	1	14.919	14.9191	31.58	0.000
X5	1	6.258	6.2577	13.24	0.000
X9	1	8.794	8.7939	18.61	0.000
X10	1	5.194	5.1939	10.99	0.001
X11	1	11.791	11.7906	24.96	0.000
X2*X2	1	2.415	2.4145	5.11	0.024
X9*X9	1	8.502	8.5021	18.00	0.000
X10*X10	1	2.075	2.0747	4.39	0.036
X2*X11	1	7.676	7.6763	16.25	0.000
X5*X11	1	6.906	6.9059	14.62	0.000
Error	985	465.377	0.4725		
Lack-of-Fit	840	465.377	0.5540	*	*
Pure Error	145	0.000	0.0000		
Total	996	701.729			

#### Model Summary

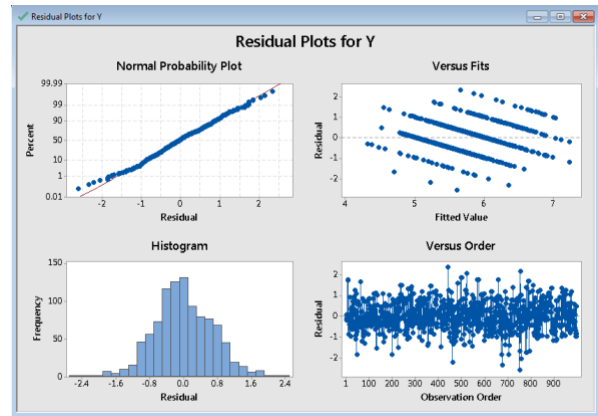
S	R-sq	R-sq(adj)	R-sq(pred)
0.687360	33.58%	32.94%	31.65%

#### Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-26.52	6.60	-4.02	0.000	
X1	0.0427	0.0177	2.41	0.016	1.94
X2	-7.52	1.34	-5.62	0.000	136.05
X5	25.79	7.09	3.64	0.000	277.54
X9	17.59	4.08	4.31	0.000	853.17
X10	1.431	0.432	3.32	0.001	15.27
X11	0.2887	0.0578	5.00	0.000	7.67
X2*X2	0.994	0.440	2.26	0.024	15.82
X9*X9	-2.641	0.623	-4.24	0.000	852.78
X10*X10	-0.498	0.238	-2.10	0.036	14.01
X2*X11	0.521	0.129	4.03	0.000	133.79
X5*X11	-2.818	0.737	-3.82	0.000	274.08

#### Regression Equation

Y = -26.52 + 0.0427 X1 - 7.52 X2 + 25.79 X5 + 17.59 X9 + 1.431 X10 + 0.2887 X11 + 0.994 X2\*X2 - 2.641 X9\*X9 - 0.498 X10\*X10 + 0.521 X2\*X11 - 2.818 X5\*X11



#### Regression Equation

$$Y = -26.52 + 0.0427 X1 - 7.52 X2 + 25.79 X5 + 17.59 X9 + 1.431 X10 + 0.2887 X11 + 0.994 X2^2 - 2.641 X9^2 - 0.498 X10^2 + 0.521 X2 X11 - 2.818 X5 X11$$

Model SSE: **465.377**

Note: This model does have many of the VIF values above 10, which indicates multi collinearity. Yet, a prediction was tried using this model on the test data.

The outcome was with an SSE: **580.278**

## STANDARDIZED Variables:

### Model 2:

#### Regression Analysis: Y versus S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, S11

Backward Elimination of Terms

$\alpha$  to remove = 0.1

#### Analysis of Variance

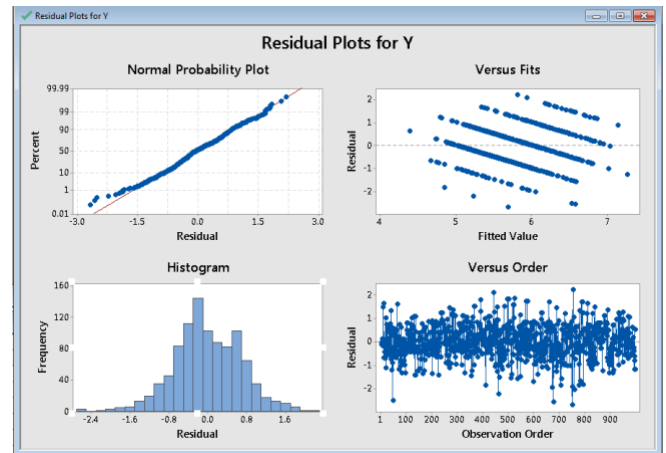
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	10	223.264	22.3264	46.01	0.000
S1	1	6.354	6.3544	13.09	0.000
S2	1	21.884	21.8845	45.10	0.000
S4	1	3.359	3.3594	6.92	0.009
S5	1	2.798	2.7979	5.77	0.017
S6	1	2.478	2.4779	5.11	0.024
S7	1	3.705	3.7050	7.64	0.006
S8	1	5.741	5.7406	11.83	0.001
S9	1	5.086	5.0863	10.48	0.001
S10	1	9.741	9.7406	20.07	0.000
S11	1	19.032	19.0322	39.22	0.000
Error	986	478.465	0.4853		
Lack-of-Fit	842	478.465	0.5682	*	*
Pure Error	144	0.000	0.0000		
Total	996	701.729			

#### Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.696605	31.82%	31.12%	30.05%

#### Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	5.6991	0.0221	258.33	0.000	
S1	0.2262	0.0625	3.62	0.000	8.02
S2	-0.2002	0.0298	-6.72	0.000	1.82
S4	0.1439	0.0547	2.63	0.009	6.14
S5	-0.0694	0.0289	-2.40	0.017	1.71
S6	0.0909	0.0402	2.26	0.024	3.32
S7	-0.1315	0.0476	-2.76	0.006	4.65
S8	-0.2674	0.0777	-3.44	0.001	12.40
S9	0.1239	0.0383	3.24	0.001	3.01
S10	0.1384	0.0309	4.48	0.000	1.96
S11	0.2407	0.0384	6.26	0.000	3.03



### Observed outliers:

-1.84342	0.56964	-0.96608	-0.57651	-0.3320	-1.02660	-0.54894	-0.93905	4.16028	-0.16398	2.86113	-2.51979
-0.32291	-0.48374	0.84483	0.10930	-0.5273	-0.40630	-0.54894	-1.82476	-1.41601	-0.92464	2.38187	-2.56837
0.43735	-0.79976	-0.62653	2.64683	0.0587	0.83429	1.58096	0.75856	-1.48011	-0.46825	-0.30198	-2.67906

Observed high VIFs : For S8 the VIF value is 12.4. Hence dropping this variable from the model.

## Regression Analysis: Y versus S1, S2, S3, S4, S5, S6, S7, S9, S10, S11

Backward Elimination of Terms

α to remove = 0.1

Analysis of Variance

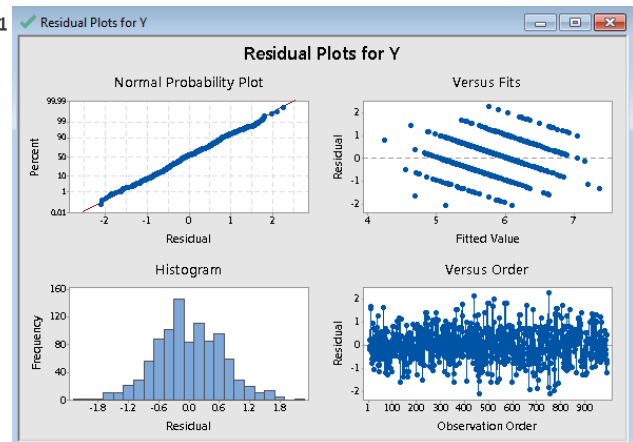
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	6	223.629	37.271	79.11	0.000
S2	1	42.888	42.888	91.03	0.000
S5	1	5.343	5.343	11.34	0.001
S6	1	1.784	1.784	3.79	0.052
S7	1	4.567	4.567	9.69	0.002
S10	1	6.896	6.896	14.64	0.000
S11	1	114.596	114.596	243.24	0.000
Error	987	465.004	0.471		
Lack-of-Fit	843	465.004	0.552	*	*
Pure Error	144	0.000	0.000		
Total	993	688.633			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.686388	32.47%	32.06%	31.44%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	5.7070	0.0218	262.14	0.000	
S2	-0.2399	0.0251	-9.54	0.000	1.34
S5	-0.0915	0.0272	-3.37	0.001	1.56
S6	0.0759	0.0390	1.95	0.052	3.21
S7	-0.1272	0.0408	-3.11	0.002	3.52
S10	0.1023	0.0268	3.83	0.000	1.51
S11	0.3644	0.0234	15.60	0.000	1.14



Regression Equation

$$\begin{aligned} Y = & 5.7070 - 0.2399 S2 - 0.0915 S5 \\ & + 0.0759 S6 - 0.1272 S7 + 0.1023 S10 \\ & + 0.3644 S11 \end{aligned}$$

SSE: 465

All VIF values are good (below 10)

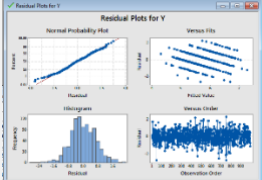
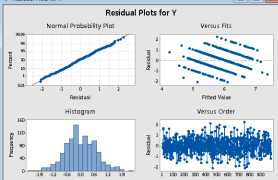
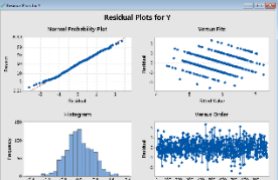
Applying Prediction on the standardized Test data set.

SSE: 584.7235

TABLE 2 : SIX STAGES

Stages	Non Standardized Model - 1	Standardized Variable Model-2	Non Standardized Quadratic Model 1_A
1 (Selection criteria)	Based on identified correlation (X6, X7) and (X1, X8). The variable X7 was dropped initially. Y → Quality of wine	S8 was dropped from model due to high VIF value, indicating multi collinearity Y → Quality of wine	Same as Model 1
2 (Variable Transformation)	Two outliers were noted (row nos. 753 and 945)	Three outliers identified and dropped (751, 676 & 46)	Same as Model 1



<b>3 (Assumption Check)</b>	 <p>Normality: ok Constant variance: ok Independence: Yes</p>	 <p>Normality: ok Constant variance: ok Independence: Yes</p>	 <p>Normality: ok Constant variance: ok Independence: Yes</p>
<b>4 (Variable Selection)</b>	<ul style="list-style-type: none"> <li>Due to insignificance, X6 was dropped</li> <li>Stepwise – backward Elimination method</li> </ul>	<ul style="list-style-type: none"> <li>Due to insignificance, S6 was dropped</li> <li>Stepwise – backward Elimination method</li> </ul>	<ul style="list-style-type: none"> <li>Same as Model 1</li> </ul>
<b>5 (Analysis Terms)</b>	<p>Regression Equation</p> $Y = 1.080 + 0.0389 X1 - 1.283 X2 - 1.469 X5 + 0.394 X9 + 0.471 X10 + 0.3352 X11$ <p>SSE : <b>487.096</b> Predicted SSE: <b>587.6833</b> VIF : under 10 (no multicollinearity)</p>	<p>Regression Equation</p> $Y = 5.7070 - 0.2399 S2 - 0.0915 S5 + 0.0759 S6 - 0.1272 S7 + 0.1023 S10 + 0.3644 S11$ <p><b>SSE: 465</b> <b>Predict SSE: 584.7235</b> VIF : good</p>	<p>Regression Equation</p> $Y = -26.52 + 0.0427 X1 - 7.52 X2 + 25.79 X5 + 17.59 X9 + 1.431 X10 + 0.2887 X11 + 0.994 X2*X2 - 2.641 X9*X9 - 0.498 X10*X10 + 0.521 X2*X11 - 2.818 X5*X11$ <p><b>SSE: 465.377</b> <b>Predict SSE: 580.278</b> <b>VIF ( showing high values, multi collinearity)</b></p>
<b>6 (Statistical Significant)</b>	R-sq : 30.59	R-sq: 32.47	R-Sq : 31.82
<b>Observations</b>	<ul style="list-style-type: none"> <li>Highest SSE</li> <li>VIF good</li> </ul>	<ul style="list-style-type: none"> <li>Better SSE</li> <li>VIF good</li> <li><b>BEST MODEL of 3, compared to Model 1.</b></li> </ul>	<ul style="list-style-type: none"> <li>Lowest SSE</li> <li>VIF bad (not suitable, even though lowest SSE)</li> </ul>

B3:

Perform Regression Analysis using all the 2000 observations to find the best predictor model for the quality of wine. Comment and show what you have done on each of the six stages of regression analysis and why. Compare and contrast it with the B2 model.

Solution:

The train dataset has all the datasets – 2000 observations

### Non Standardized model 1:

A backward elimination approach was adopted with the all the variables

### Regression Analysis: Y versus X1, X2, X4, X5, X6, X7, X10, X11

Backward Elimination of Terms

Candidate terms: X1, X2, X4, X5, X6, X7, X10, X11

	-----Step 1-----		-----Step 2-----		-----Step 3-----	
	Coef	P	Coef	P	Coef	P
Constant	2.759		2.787		2.716	
X1	-0.0087	0.450	-0.0078	0.492		
X2	-1.366	0.000	-1.370	0.000	-1.365	0.000
X4	0.00313	0.527				
X5	-1.953	0.000	-1.973	0.000	-1.984	0.000
X6	0.00590	0.000	0.00604	0.000	0.00605	0.000
X7	-0.001944	0.000	-0.001866	0.000	-0.001773	0.000
X10	0.701	0.000	0.697	0.000	0.680	0.000
X11	0.3281	0.000	0.3255	0.000	0.3264	0.000
S		0.725007		0.724898		0.724802
R-sq		30.85%		30.83%		30.81%
R-sq(adj)		30.57%		30.59%		30.61%
R-sq(pred)		30.12%		30.19%		30.26%
Mallows' Cp		9.00		7.40		5.87

$\alpha$  to remove = 0.1

# Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	6	465.16	77.526	147.57	0.000
X2	1	97.03	97.034	184.71	0.000
X5	1	10.40	10.401	19.80	0.000
X6	1	7.74	7.735	14.72	0.000
X7	1	7.46	7.462	14.20	0.000
X10	1	20.56	20.563	39.14	0.000
X11	1	199.62	199.622	379.99	0.000
Error	1988	1044.37	0.525		
Lack-of-Fit	1680	1044.37	0.622	*	*
Pure Error	308	0.00	0.000		
Total	1994	1509.53			

## Model Summary

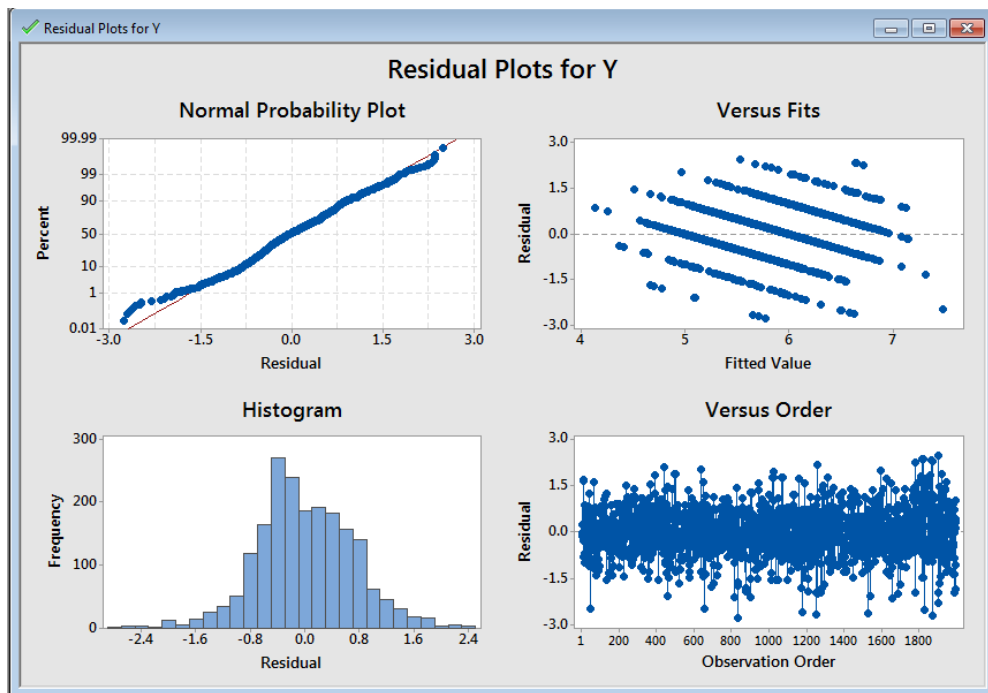
S	R-sq	R-sq(adj)	R-sq(pred)
0.724802	30.81%	30.61%	30.26%

## Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2.716	0.208	13.05	0.000	
X2	-1.365	0.100	-13.59	0.000	1.37
X5	-1.984	0.446	-4.45	0.000	1.51
X6	0.00605	0.00158	3.84	0.000	2.85
X7	-0.001773	0.000471	-3.77	0.000	3.28
X10	0.680	0.109	6.26	0.000	1.36
X11	0.3264	0.0167	19.49	0.000	1.19

## Regression Equation

$$Y = 2.716 - 1.365 X2 - 1.984 X5 + 0.00605 X6 - 0.001773 X7 + 0.680 X10 + 0.3264 X11$$



SSE - 1044.37

Standardized Model :

Outliers :

1253	-1.17546	-0.87655	0.72498	-0.25054	-0.8897	-1.18766	0.19592	-1.70304	1.72271	-0.87132	1.42453	-3.25704	-3.25035
1445	-0.40741	-0.45330	0.02041	1.51537	-0.6883	-0.55506	-0.49300	-0.81940	-0.09098	-1.04331	1.23527	-3.20288	-3.18696
1740	-0.52557	-0.08296	0.49012	0.00846	-1.0464	-1.18766	-1.24600	-1.62941	0.34681	-1.21529	2.27618	-3.47899	-3.46740
1774	0.77419	-0.71783	0.78369	1.42119	-0.7555	0.13504	0.43624	0.39562	-0.34114	-0.69934	0.19437	3.07488	3.07480

Model 3 : ( Post removing the above Outliers)

## Regression Analysis: Y versus A2, A5, A6, A7, A10, A11

### Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	6	465.16	77.526	147.57	0.000
A2	1	97.03	97.034	184.71	0.000
A5	1	10.40	10.401	19.80	0.000
A6	1	7.74	7.735	14.72	0.000
A7	1	7.46	7.462	14.20	0.000
A10	1	20.56	20.563	39.14	0.000
A11	1	199.62	199.622	379.99	0.000
Error	1988	1044.37	0.525		
Lack-of-Fit	1675	1044.37	0.624	*	*
Pure Error	313	0.00	0.000		
Total	1994	1509.53			

### Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.724802	30.81%	30.61%	30.26%

---

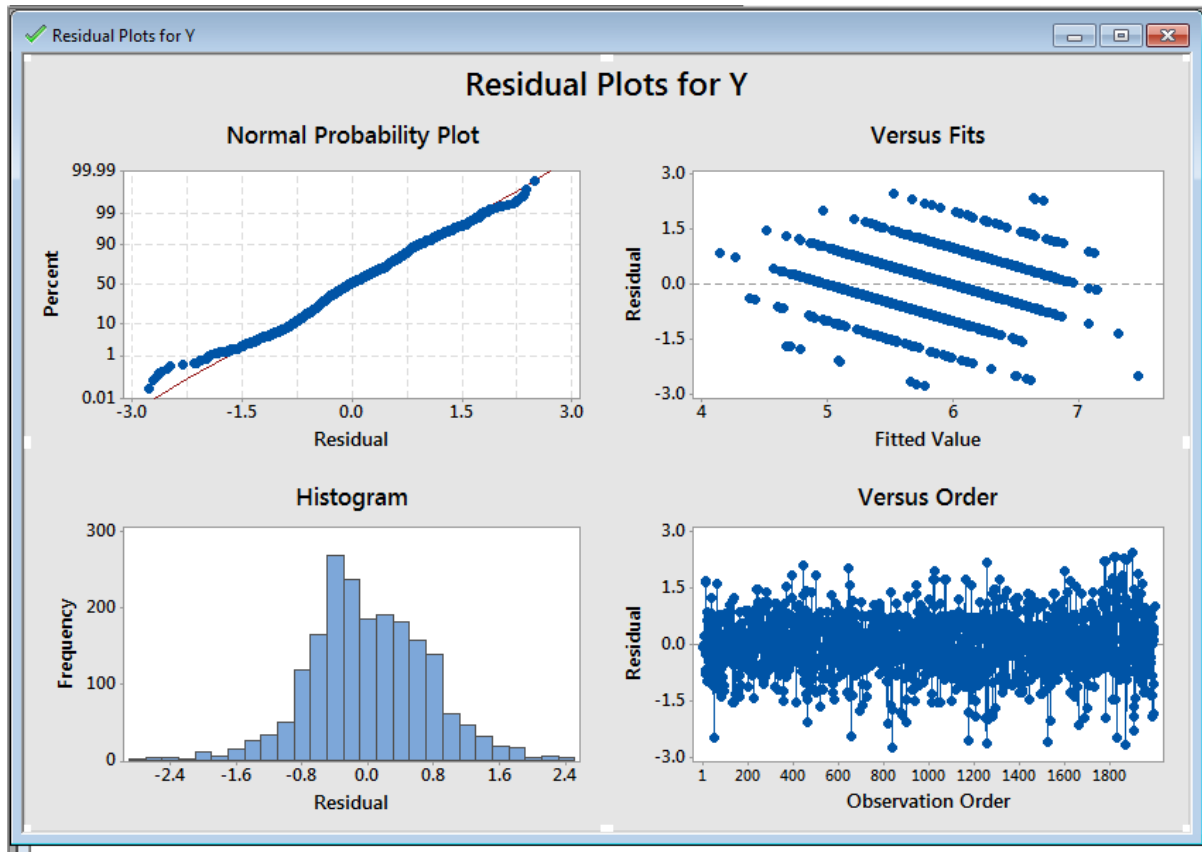
### Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	5.7328	0.0162	353.28	0.000	
A2	-0.2581	0.0190	-13.59	0.000	1.37
A5	-0.0887	0.0199	-4.45	0.000	1.51
A6	0.1052	0.0274	3.84	0.000	2.85
A7	-0.1107	0.0294	-3.77	0.000	3.28
A10	0.1186	0.0190	6.26	0.000	1.36
A11	0.3450	0.0177	19.49	0.000	1.19

---

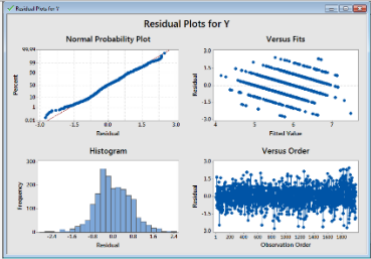
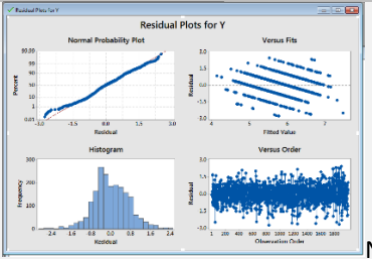
### Regression Equation

$$Y = 5.7328 - 0.2581 A2 - 0.0887 A5 + 0.1052 A6 - 0.1107 A7 + 0.1186 A10 + 0.3450 A11$$



SSE (Post Standardizing) - 1044.37

**TABLE 3 : SIX STAGES**

Stages	Non Standardized	Standardized
<b>1 (Selection criteria)</b>	All the variables were taken in to consideration. Y → Quality of Wine	Similar approach as non standardized
<b>2 (Variable Transformation)</b>	Three outliers were noted with residual values between -3.2 to -3.5.	4 outliers were noted and removed
<b>3 (Assumption Check)</b>	 <p>Normality: ok Constant variance: ok Independence: Yes</p>	 <p>Normality: ok Constant variance: ok Independence: Yes</p>

<b>4 (Variable Selection)</b>	<ul style="list-style-type: none"> <li>• X3 dropped due to high P value</li> <li>• X8 dropped due to high VIF value</li> <li>• X9 dropped due to high P value</li> <li>• Stepwise – backward Elimination method , X1 dropped</li> </ul>	<ul style="list-style-type: none"> <li>• Due to insignificance, A1, A4 was dropped</li> </ul>
<b>5 (Analysis Terms)</b>	Regression Equation $Y = 2.716 - 1.365 X2 - 1.984 X5 + 0.00605 X6 - 0.001773 X7 + 0.680 X10 + 0.3264 X11$ <b>SSE - 1044.37</b> VIF : under 10 (no multi collinearity)	Regression Equation $Y = 5.7328 - 0.2581 A2 - 0.0887 A5 + 0.1052 A6 - 0.1107 A7 + 0.1186 A10 + 0.3450 A11$ <b>SSE : 1044.37</b> VIF : under 10 (no multicollinearity)
<b>6 (Statistical Significant)</b>	<b>R-Sq : 30.81</b> <b>Adj R.Sq : 30.61</b>	<b>R-Sq : 30.81</b> <b>Adj R.Sq : 30.61</b>
<b>Observations</b>	No difference between models	

### Comparing with B2

<b>B3 standardized</b>	<b>B2 Standardized</b>
<b>Regression Equation</b> $Y = 5.7328 - 0.2581 A2 - 0.0887 A5 + 0.1052 A6 - 0.1107 A7 + 0.1186 A10 + 0.3450 A11$ <b>SSE : 1044.37</b> <b>VIF : under 10 (no multicollinearity)</b>	<b>Regression Equation</b> $Y = 5.7070 - 0.2399 S2 - 0.0915 S5 + 0.0759 S6 - 0.1272 S7 + 0.1023 S10 + 0.3644 S11$ <b>SSE: 465</b> <b>Predict SSE: 584.7235</b> VIF : good
R-Sq : 30.81 Adj R.Sq : 30.61	R-sq: 32.47

### Findings:

B2 SSE is low, as the number of points are low.

B3 has lower R-Square indicating better model.

B4:

Perform Regression Analysis using observations 1-1000 (1000 Red wine data) to find the best predictor model for the quality of wine. Test the Comment and show what you have done on each of the six stages of regression analysis and why

Solution:

Model 1: Non standardized variables

**Variables that were removed during iterations: X1, X3,X4,X6,X8,X9**

**Post removing Insignificant variables:**

### Regression Analysis: Y versus X2, X5, X7, X10, X11

#### Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	221.332	44.2663	105.75	0.000
X2	1	28.688	28.6876	68.53	0.000
X5	1	4.586	4.5857	10.95	0.001
X7	1	13.992	13.9921	33.43	0.000
X10	1	13.067	13.0665	31.22	0.000
X11	1	84.057	84.0574	200.81	0.000
Error	993	415.667	0.4186		
Lack-of-Fit	843	415.667	0.4931	*	*
Pure Error	150	0.000	0.0000		
Total	998	636.999			

#### Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.646991	34.75%	34.42%	33.80%

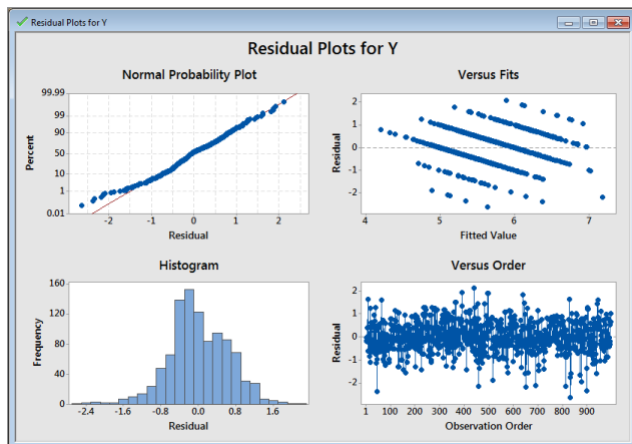
#### Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2.899	0.249	11.64	0.000	
X2	-1.001	0.121	-8.28	0.000	1.11
X5	-1.527	0.461	-3.31	0.001	1.27
X7	-0.003654	0.000632	-5.78	0.000	1.06
X10	0.719	0.129	5.59	0.000	1.33
X11	0.2986	0.0211	14.17	0.000	1.12

#### Regression Equation

$$Y = 2.899 - 1.001 X2 - 1.527 X5 - 0.003654 X7 + 0.719 X10 + 0.2986 X11$$





SSE - 415.6673

Model 2: Standardized

## Regression Analysis: Y versus A2, A5, A7, A10, A11

### Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	221.332	44.2663	105.75	0.000
A2	1	28.688	28.6876	68.53	0.000
A5	1	4.586	4.5857	10.95	0.001
A7	1	13.992	13.9921	33.43	0.000
A10	1	13.067	13.0665	31.22	0.000
A11	1	84.057	84.0574	200.81	0.000
Error	993	415.667	0.4186		
Lack-of-Fit	843	415.667	0.4931	*	*
Pure Error	150	0.000	0.0000		
Total	998	636.999			

### Model Summary

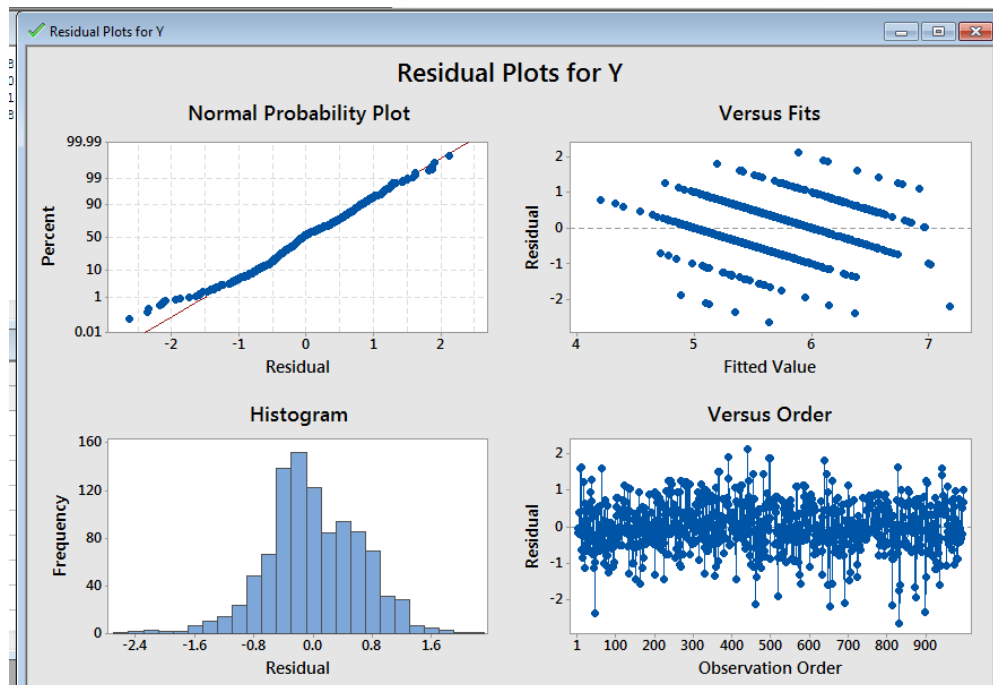
S	R-sq	R-sq(adj)	R-sq(pred)
0.646991	34.75%	34.42%	33.80%

### Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	5.5936	0.0205	273.26	0.000	
A2	-0.1787	0.0216	-8.28	0.000	1.11
A5	-0.0762	0.0230	-3.31	0.001	1.27
A7	-0.1217	0.0211	-5.78	0.000	1.06
A10	0.1317	0.0236	5.59	0.000	1.33
A11	0.3076	0.0217	14.17	0.000	1.12

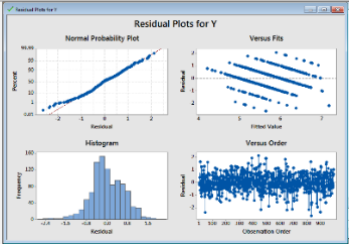
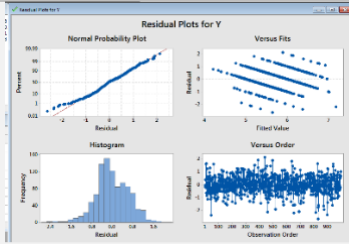
### Regression Equation

$$Y = 5.5936 - 0.1787 A2 - 0.0762 A5 - 0.1217 A7 + 0.1317 A10 + 0.3076 A11$$



SSE (Post Standardizing) - 415.6673

TABLE 4 : SIX STAGES

Stages	Non Standardized Model	Standardized
<b>1 (Selection criteria)</b>	All the variables were taken in to consideration. Y → Quality of Wine	All standardized variables were taken in to consideration. Y → Quality of wine
<b>2 (Variable Transformation)</b>	No outliers identified.	No outliers identified.
<b>3 (Assumption Check)</b>	 <p>Normality: ok Constant variance: ok Independence: Yes</p>	 <p>Normality: ok Constant variance: ok Independence: Yes</p>

<b>4 (Variable Selection)</b>	<ul style="list-style-type: none"> <li><b>X1, X3,X4,X6,X8,X9 were dropped based on P values at each step in backward elimination</b></li> </ul>	<ul style="list-style-type: none"> <li><b>A1, A3,A4,A6,A8,A9 Variables were dropped in same fashion</b></li> </ul>
<b>5 (Analysis Terms)</b>	Regression Equation $Y = 2.899 - 1.001 X2 - 1.527 X5 + 0.003654 X7 + 0.719 X10 + 0.2986 X11$ <b>SSE : 415.667</b> <b>VIF : under 10 (no multi collinearity)</b>	Regression Equation $Y = 5.5936 - 0.1787 A2 - 0.0762 A5 - 0.1217 A7 + 0.1317 A10 + 0.3076 A11$ <b>SSE : 415.6673</b> <b>VIF : under 10 (no multicollinearity)</b>
<b>6 (Statistical Significant)</b>	<b>R-Sq : 34.75</b> <b>Adj R.Sq : 34.42</b>	<b>R-Sq : 34.75</b> <b>Adj R.Sq : 34.42</b>
<b>Observations</b>	Both the models (standardized as well as non-standardized) for the red wine data has similar statistical significance.	

## B5:

Perform Regression Analysis using observations 1001 - 2000 (1000 White wine data) to find the best predictor model for the quality of wine. Comment and show what you have done on each of the six stages of regression analysis and why. Compare it with the model of B4 Red wine data

Solution:

Final Model after few iterations as follows. The details are discussed in the table.

Model 1 :

During Nonstandardized variable model iteration, Identified Outliers during the process are

252	8.5	0.260	0.21	16.20	0.074	41.0	197.0	0.99800	3.02	0.50	9.8	3
254	5.8	0.240	0.44	3.50	0.029	5.0	109.0	0.99130	3.53	0.43	11.7	3
446	7.1	0.320	0.32	11.00	0.038	16.0	66.0	0.99370	3.24	0.40	11.5	3
741	6.9	0.390	0.40	4.60	0.022	5.0	19.0	0.99150	3.31	0.37	12.6	3
775	9.1	0.270	0.45	10.60	0.035	28.0	124.0	0.99700	3.20	0.46	10.4	9
905	6.9	0.210	0.28	2.40	0.056	49.0	159.0	0.99440	3.02	0.47	8.8	8

**Post removing Outliers:**

### Regression Analysis: Y versus X1, X2, X4, X6, X8, X9, X10

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	7	261.155	37.308	65.65	0.000
X1	1	24.870	24.870	43.76	0.000
X2	1	24.354	24.354	42.85	0.000
X4	1	70.133	70.133	123.40	0.000
X6	1	4.217	4.217	7.42	0.007
X8	1	135.266	135.266	238.01	0.000
X9	1	47.662	47.662	83.86	0.000
X10	1	22.241	22.241	39.14	0.000
Error	986	560.362	0.568		
Lack-of-Fit	827	560.362	0.678	*	*
Pure Error	159	0.000	0.000		
Total	993	821.517			

Model Summary

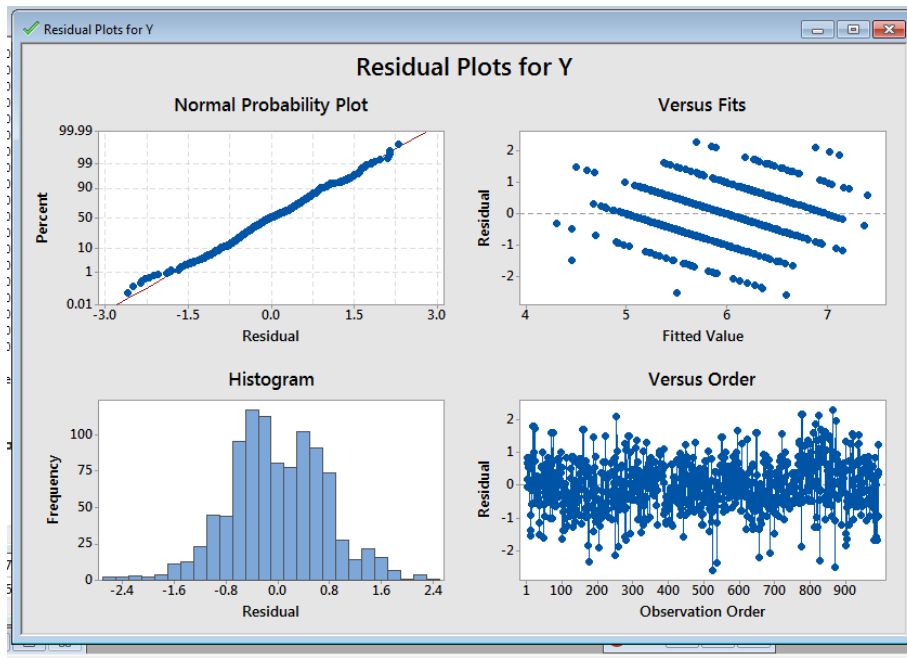
S	R-sq	R-sq(adj)	R-sq(pred)
0.753869	31.79%	31.31%	30.45%

### Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	320.1	20.7	15.49	0.000	
X1	0.2598	0.0393	6.62	0.000	1.57
X2	-1.655	0.253	-6.55	0.000	1.08
X4	0.1224	0.0110	11.11	0.000	5.66
X6	0.00436	0.00160	2.72	0.007	1.28
X8	-324.8	21.1	-15.43	0.000	5.84
X9	1.814	0.198	9.16	0.000	1.52
X10	1.414	0.226	6.26	0.000	1.10

### Regression Equation

$$Y = 320.1 + 0.2598 X1 - 1.655 X2 + 0.1224 X4 + 0.00436 X6 - 324.8 X8 + 1.814 X9 + 1.414 X10$$



**SSE - 560.3621**

Model 2: Standardized variables

Identified outliers during the iterations

252	2.15001	-0.23804	-0.9798	1.86923	1.1284	0.28886	1.16004	1.27574	-1.27439	0.04140	-0.32476	-3.01854
254	-1.36584	-0.44188	0.7787	-0.58974	-0.7615	-1.83949	-0.80803	-1.16444	2.14257	-0.58991	1.43222	-3.59321
446	0.32697	0.37349	-0.1388	0.86241	-0.3836	-1.18916	-1.76970	-0.29034	0.19959	-0.86047	1.24727	-3.39752

741	0.06654	1.08695	0.4729	-0.37675	-1.0555	-1.83949	-2.82082	-1.09160	0.66859	-1.13103	2.26447	-3.21328
-----	---------	---------	--------	----------	---------	----------	----------	----------	---------	----------	---------	----------

775	2.93131	-0.13612	0.8552	0.78496	-0.5096	-0.47971	-0.47256	0.91153	-0.06841	-0.31935	0.23007	2.97985
-----	---------	----------	--------	---------	---------	----------	----------	---------	----------	----------	---------	---------

905	0.06654	-0.74765	-0.4446	-0.80272	0.3724	0.76183	0.31019	-0.03540	-1.27439	-0.22916	-1.24949	2.70864
-----	---------	----------	---------	----------	--------	---------	---------	----------	----------	----------	----------	---------

## Regression Analysis: Y versus A1, A2, A4, A6, A8, A9, A10

### Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	7	261.155	37.308	65.65	0.000
A1	1	24.870	24.870	43.76	0.000
A2	1	24.354	24.354	42.85	0.000
A4	1	70.133	70.133	123.40	0.000
A6	1	4.217	4.217	7.42	0.007
A8	1	135.266	135.266	238.01	0.000
A9	1	47.662	47.662	83.86	0.000
A10	1	22.241	22.241	39.14	0.000
Error	986	560.362	0.568		
Lack-of-Fit	827	560.362	0.678	*	*
Pure Error	159	0.000	0.000		
Total	993	821.517			

### Model Summary

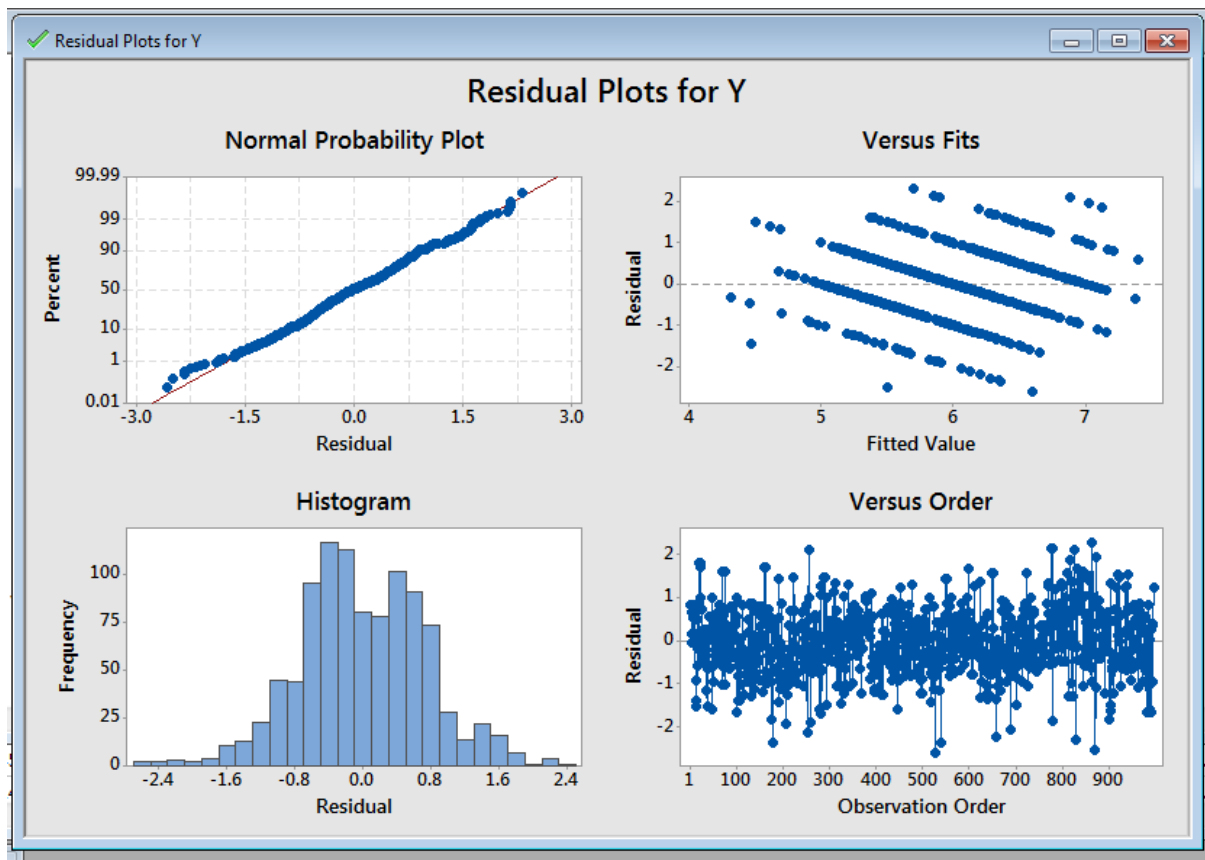
S	R-sq	R-sq(adj)	R-sq(pred)
0.753869	31.79%	31.31%	30.45%

### Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	5.8728	0.0239	245.60	0.000	
A1	0.1995	0.0302	6.62	0.000	1.57
A2	-0.1624	0.0248	-6.55	0.000	1.08
A4	0.6321	0.0569	11.11	0.000	5.66
A6	0.0738	0.0271	2.72	0.007	1.28
A8	-0.8919	0.0578	-15.43	0.000	5.84
A9	0.2707	0.0296	9.16	0.000	1.52
A10	0.1567	0.0251	6.26	0.000	1.10

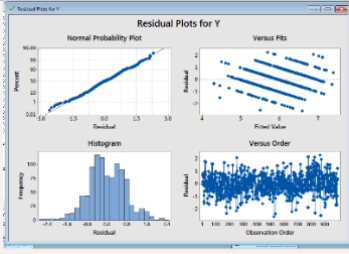
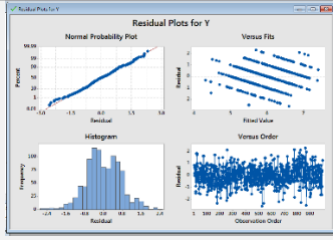
### Regression Equation:

$$Y = 5.8728 + 0.1995 A1 - 0.1624 A2 + 0.6321 A4 + 0.0738 A6 - 0.8919 A8 + 0.2707 A9 + 0.1567 A10$$



SSE ( Post Standardizing) - 560.3621

TABLE 5 : SIX STAGES

Stages	Non Standardized Model - 1	Standardized Model 2
<b>1 (Selection criteria)</b>	All the variables were taken in to consideration. Y → Quality of Wine	All variables as in Model 1 Y → Quality of wine
<b>2 (Variable Transformation)</b>	6 outliers were noted with residual values below -3	6 outliers were noted
<b>3 (Assumption Check)</b>	 <p>Normality: ok Constant variance: ok Independence: Yes</p>	 <p>Normality: ok Constant variance: ok Independence: Yes</p>
<b>4 (Variable Selection)</b>	<ul style="list-style-type: none"> <li>• X3 dropped due to high P value</li> <li>• X5 dropped due to high P value</li> <li>• X7 dropped due to high P value</li> <li>• X11 dropped due to high P value</li> </ul>	<ul style="list-style-type: none"> <li>• Variables were selected based on the understanding in model 1</li> </ul>
<b>5 (Analysis Terms)</b>	<p>Regression Equation</p> $Y = 320.1 + 0.2598 X1 - 1.655 X2 + 0.1224 X4 + 0.00436 X6 - 324.8 X8 + 1.814 X9 + 1.434 X10$ <p><b>SSE - 560.3621</b> VIF : under 10 (no multi collinearity)</p>	<p>Regression Equation</p> $Y = 5.8728 + 0.1995 A1 - 0.1624 A2 + 0.6321 A4 + 0.0738 A6 - 0.8919 A8 + 0.2707 A9 + 0.1567 A10$ <p><b>SSE : 560.3621</b> VIF : under 10 (no multi collinearity)</p>
<b>6 (Statistical Significant)</b>	<b>R-Sq : 31.79</b> <b>Adj R.Sq : 31.31</b>	<b>R-sq : 31.79</b> <b>Adj R.Sq : 31.31</b>
<b>Observations</b>	In B5, both the models show similar statistical significance, indicating no difference between the models.	

Standardized Model from B4:

<p><b>Regression Equation</b></p> $Y = 5.5936 - 0.1787 A2 - 0.0762 A5 - 0.1217 A7 + 0.1317 A10 + 0.3076 A11$ <p><b>SSE : 415.6673</b> <b>VIF : under 10 (no multicollinearity)</b></p> <p><b>R-Sq : 34.75</b> <b>Adj R.Sq : 34.42</b></p>
---

Comparison between B4 &amp; B5 models.

B4	B5
----	----



Variables influencing RED wine (B4): X2, X5, X7, X10, X11	Variables influencing WHITE wine (B5): X1, X2, X4, X6, X8. X9, X10
SSE : <b>415.6673</b> (comparatively a RED wine could be predicted with low residuals)	<b>SSE</b> : 560.3621 Not better than RED wine model .