Multiple Regression Model of Wine Quality

Data

Dataset:

Wine Quality – white wine

Attributes:

Fixed acidity	Total sulfur dioxide
Volatile acidity	Density
Citric acid	pH
Residual sugar	Sulphates
Free sulfur dioxide	Alcohol
Chlorides	Quality

Link:

http://archive.ics.uci.edu/ml/datasets/Wine+Quality

Citation:

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.

Correlation Coefficients

Pearson Correlation Coefficients, N = 4898 Prob > |r| under H0: Rho=0											
	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates	alcohol
quality	-0.11366	-0.19472	-0.00921	-0.09758	-0.20993	0.00816	-0.17474	-0.30712	0.09943	0.05368	0.43557
quality	<.0001	<.0001	0.5193	<.0001	<.0001	0.5681	<.0001	<.0001	<.0001	0.0002	<.0001

Weak linear correlation

Test of Normality

	Kolmogorov-Smirnov D	p value
Fixed acidity	0.066232	<0.0100
Volatile acidity	0.104513	<0.0100
Citric acid	0.11275	<0.0100
Residual sugar	0.136624	<0.0100
Chlorides	0.207263	<0.0100
Free sulfur dioxide	0.057682	<0.0100
Total sulfur dioxide	0.04465	<0.0100
Density	0.05221	<0.0100
pH	0.049269	<0.0100
Sulphates	0.08685	<0.0100
Alcohol	0.091573	<0.0100

Multiple Regression (Stepwise)

Number of Observations Read	4898
Number of Observations Used	4898

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	1082.20641	135.27580	239.73	<.0001
Error	4889	2758.78338	0.56428		
Corrected Total	4897	3840.98979			

Root MSE	0.75119	R-Square	0.2818
Dependent Mean	5.87791	Adj R-Sq	0.2806
Coeff Var	12.77985		

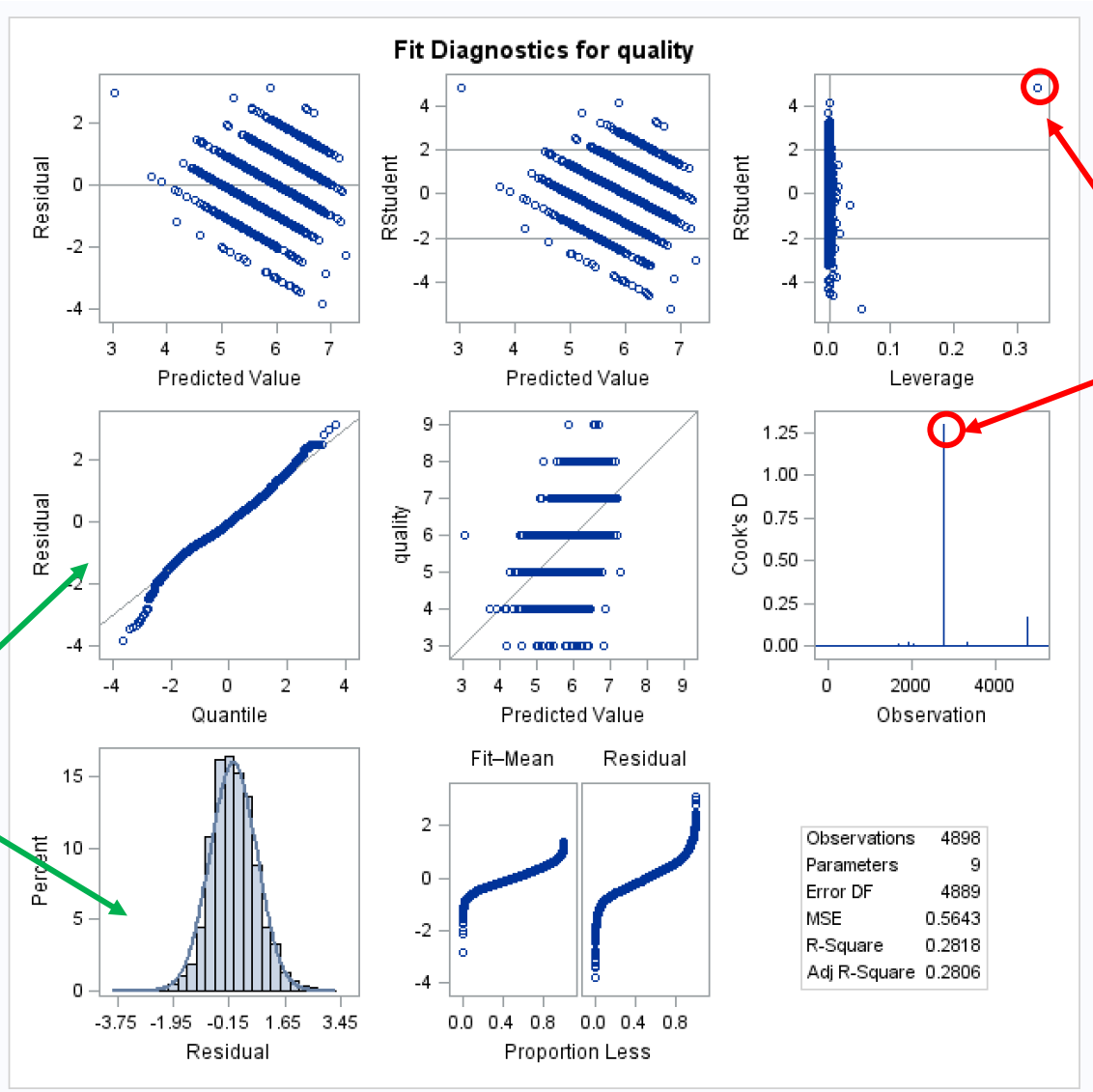
F Value: 239.73
Adj R^2: 0.2806

Parameter of Multiple Regression

Parameter Estimates											
							Heteroscedasticity Consistent				
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standard Error	t Value	Pr > |t|	Squared Partial Corr Type I	Variance Inflation
Intercept	Intercept	1	154.10624	18.10013	8.51	<.0001	41.19018	3.74	0.0002	.	0
fixed_acidity	fixed_acidity	1	0.06810	0.02043	3.33	0.0009	0.03552	1.92	0.0552	0.01292	2.57964
volatile_acidity	volatile_acidity	1	-1.88814	0.10951	-17.24	<.0001	0.11260	-16.77	<.0001	0.03946	1.05731
residual_sugar	residual_sugar	1	0.08285	0.00729	11.37	<.0001	0.01412	5.87	<.0001	0.00591	11.85425
free_sulfur_dioxide	free_sulfur_dioxide	1	0.00335	0.00067658	4.95	<.0001	0.00112	2.98	0.0029	0.00005055	1.14903
density	density	1	-154.29127	18.34398	-8.41	<.0001	41.64559	-3.70	0.0002	0.18237	26.12315
pH	pH	1	0.69421	0.10335	6.72	<.0001	0.15488	4.48	<.0001	0.04149	2.11360
sulphates	sulphates	1	0.62851	0.09997	6.29	<.0001	0.11227	5.60	<.0001	0.01481	1.12969
alcohol	alcohol	1	0.19316	0.02408	8.02	<.0001	0.05348	3.61	0.0003	0.01299	7.62284

Highly correlated

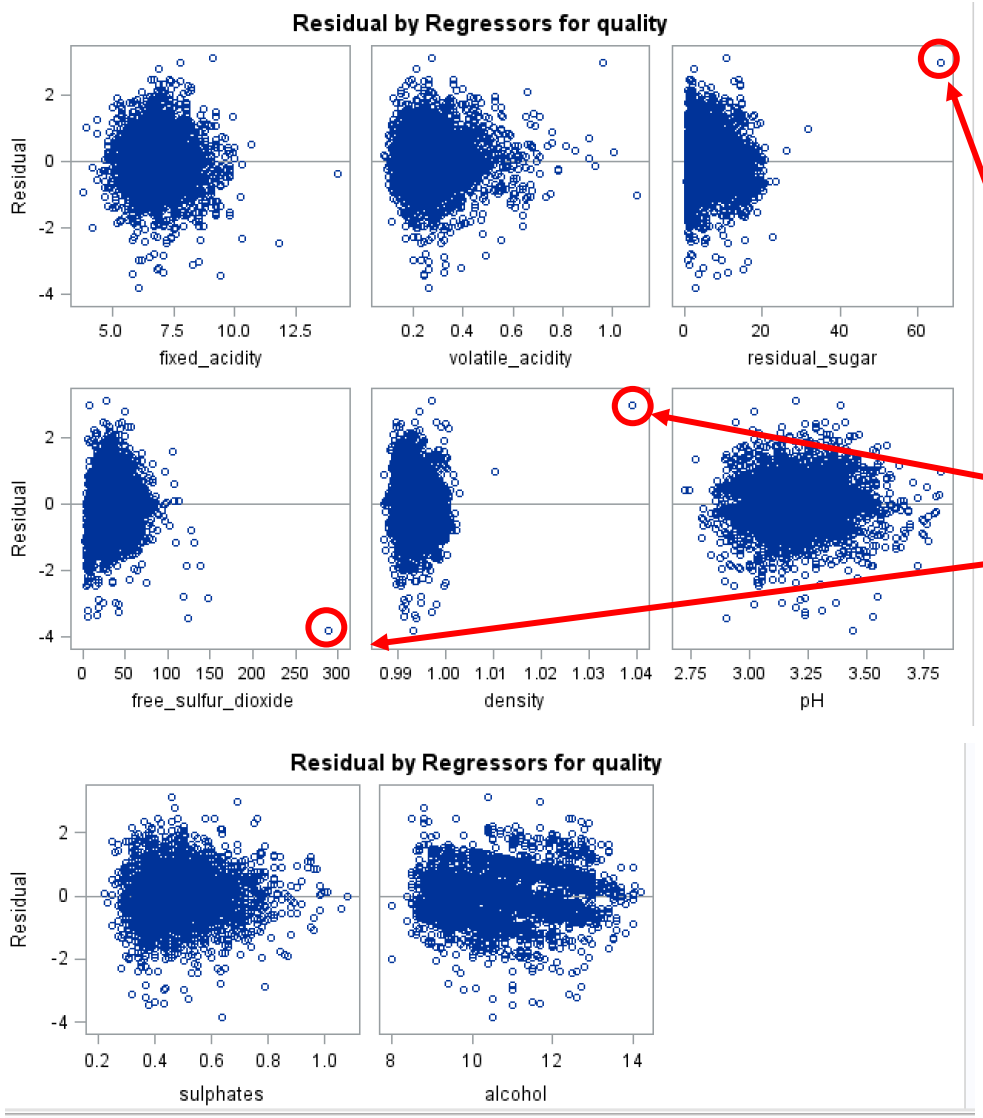
Residual



Might be high influential point

Acceptable normality

Residual



Outliers

Find High influential point

Leverage

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0.000425720	3196	0.0329547	1218
0.000428555	2135	0.0362187	1527
0.000429366	598	0.0457239	485
0.000436861	1474	0.0535734	4746
0.000438200	2838	0.3555039	2782

Cook'D

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1.27230E-11	4306	0.0122243	1418
3.35002E-11	1312	0.0178098	1932
1.49561E-10	4369	0.0202792	3308
1.88448E-10	1460	0.1419589	4746
1.94964E-10	1844	1.1343468	2782

Dffits

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
-1.25E+00	4746	0.177970	1664
-4.73E-01	3308	0.193693	18
-4.43E-01	1932	0.193693	21
-3.67E-01	1418	0.210306	775
-3.48E-01	1689	3.540237	2782

Observation #2782 is a high influential point
Remove #2782

Compare models after removing #2782

New:

Number of Observations Read	4897
Number of Observations Used	4897

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	1095.37291	136.92161	243.76	<.0001
Error	4888	2745.60197	0.56170		
Corrected Total	4896	3840.97488			

Root MSE	0.74947	R-Square	0.2852
Dependent Mean	5.87788	Adj R-Sq	0.2840
Coeff Var	12.75064		

Original:

Number of Observations Read	4898
Number of Observations Used	4898

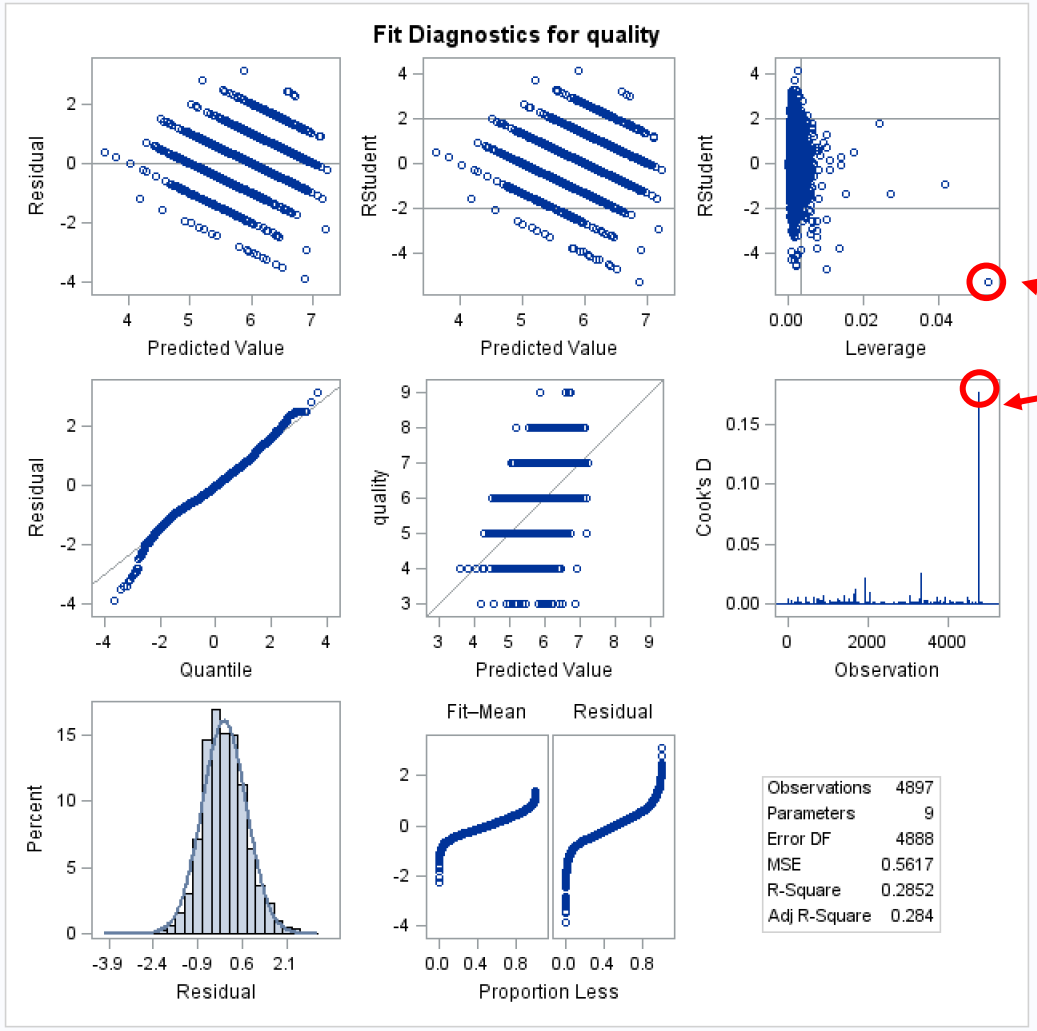
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	1082.20641	135.27580	239.73	<.0001
Error	4889	2758.78338	0.56428		
Corrected Total	4897	3840.98979			

Root MSE	0.75119	R-Square	0.2818
Dependent Mean	5.87791	Adj R-Sq	0.2806
Coeff Var	12.77985		

With higher F value and Adj. R^2

Better!

Residual of new model



Might be high influential point

Check influential point again

Leverage

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0.000428630	2135	0.0330905	1218
0.000429167	3195	0.0346025	3901
0.000441996	1647	0.0437306	1527
0.000441996	1642	0.0458430	485
0.000446825	1712	0.0536639	4745

Cook'D

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1.17203E-10	1325	0.0110364	1689
2.31558E-10	1059	0.0119163	1418
2.31558E-10	1057	0.0180039	1932
4.51494E-10	3781	0.0217179	3307
5.15762E-10	4850	0.1453987	4745

Dffits

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
-1.27E+00	4745	0.195644	18
-4.90E-01	3307	0.195644	21
-4.46E-01	1932	0.210506	775
-3.62E-01	1418	0.293130	1654
-3.49E-01	1689	0.293130	1664

No significant influence

Multicollinearity

Parameter Estimates											
							Heteroscedasticity Consistent				
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standard Error	t Value	Pr > |t|	Squared Partial Corr Type I	Variance Inflation
Intercept	Intercept	1	211.08254	21.55114	9.79	<.0001	22.83260	9.24	<.0001	.	0
fixed_acidity	fixed_acidity	1	0.11175	0.02229	5.01	<.0001	0.02424	4.61	<.0001	0.01293	3.08282
volatile_acidity	volatile_acidity	1	-1.88001	0.10927	-17.21	<.0001	0.11249	-16.71	<.0001	0.03999	1.04753
residual_sugar	residual_sugar	1	0.10140	0.00822	12.34	<.0001	0.00858	11.82	<.0001	0.00669	14.72117
free_sulfur_dioxide	free_sulfur_dioxide	1	0.00351	0.00067580	5.19	<.0001	0.00113	3.10	0.0019	0.00009265	1.15106
density	density	1	-211.85542	21.82125	-9.71	<.0001	23.13326	-9.16	<.0001	0.19283	35.42209
pH	pH	1	0.86869	0.10922	7.95	<.0001	0.11456	7.58	<.0001	0.04144	2.37060
sulphates	sulphates	1	0.70347	0.10094	6.97	<.0001	0.10222	6.88	<.0001	0.01484	1.15615
alcohol	alcohol	1	0.11941	0.02845	4.20	<.0001	0.02965	4.03	<.0001	0.00359	10.68115

Highly correlated

Remove Density

Compare models

New(Remove Density):

Number of Observations Read	4897
Number of Observations Used	4897

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	1047.70316	116.41146	203.67	<.0001
Error	4887	2793.27173	0.57157		
Corrected Total	4896	3840.97488			

Root MSE	0.75602	R-Square	0.2728
Dependent Mean	5.87788	Adj R-Sq	0.2714
Coeff Var	12.86217		

Original(Remove #2782):

Number of Observations Read	4897
Number of Observations Used	4897

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	1095.37291	136.92161	243.76	<.0001
Error	4888	2745.60197	0.56170		
Corrected Total	4896	3840.97488			

Root MSE	0.74947	R-Square	0.2852
Dependent Mean	5.87788	Adj R-Sq	0.2840
Coeff Var	12.75064		

With lower F value and Adj. R^2

Better!

Result

Estimated Regression Equation:

quality = 211.08254
+ 0.11175 * fixed acidity
- 1.88001 * volatile acidity
+ 0.10140 * residual sugar
+ 0.00351 * free sulfur dioxide
- 211.85542 * density
+ 0.86869 * pH
+ 0.70347 * sulphates
+ 0.11941 * alcohol

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	211.08254	21.55114	9.79	<.0001
fixed_acidity	fixed_acidity	1	0.11175	0.02229	5.01	<.0001
volatile_acidity	volatile_acidity	1	-1.88001	0.10927	-17.21	<.0001
residual_sugar	residual_sugar	1	0.10140	0.00822	12.34	<.0001
free_sulfur_dioxide	free_sulfur_dioxide	1	0.00351	0.00067580	5.19	<.0001
density	density	1	-211.85542	21.82125	-9.71	<.0001
pH	pH	1	0.86869	0.10922	7.95	<.0001
sulphates	sulphates	1	0.70347	0.10094	6.97	<.0001
alcohol	alcohol	1	0.11941	0.02845	4.20	<.0001