Module 8: Portfolio Project

Option 2: Build a Wine Quality Multiple Linear Regression Model Using SAS Studio

MIS 500: Foundations of Data Analytics

Dr. Osama Morad

February 7, 2022

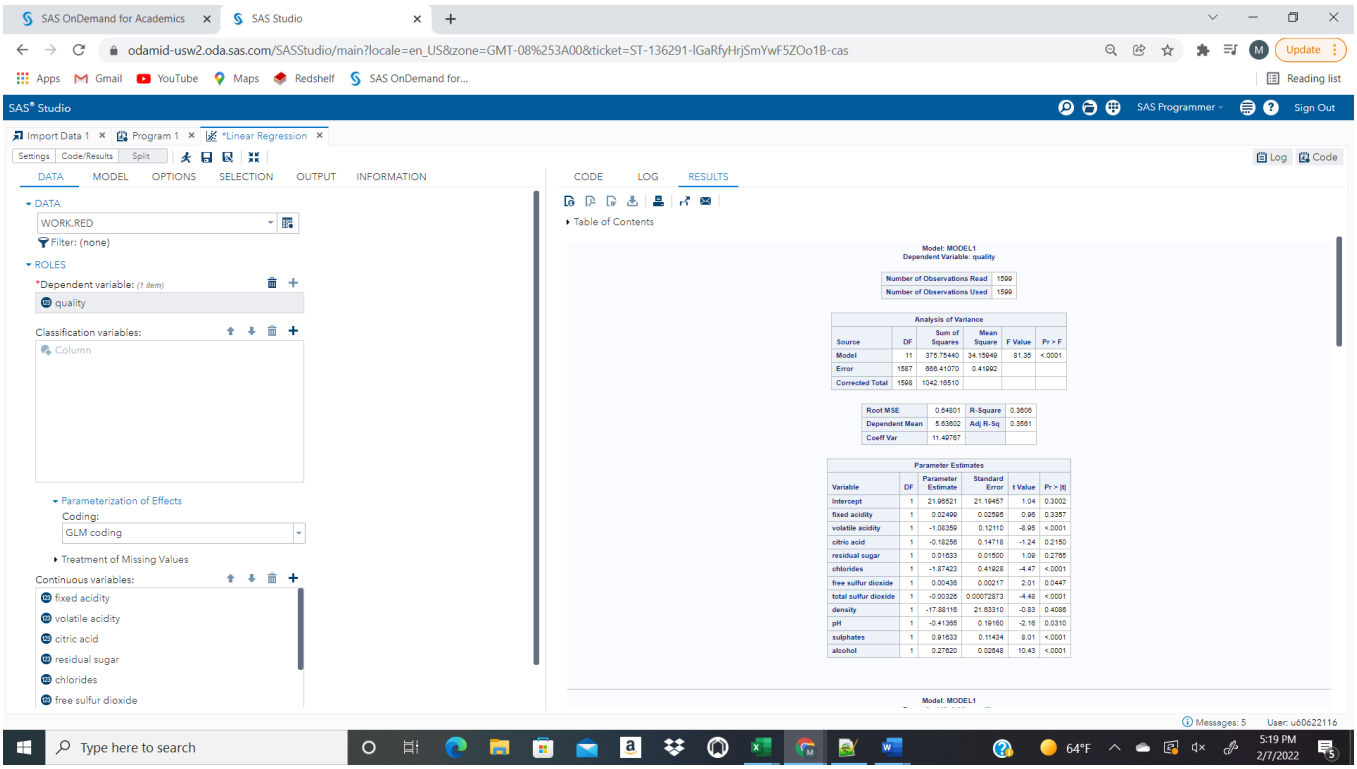
The quality of wine is subjective, but with predictive analytics it could be possible to predict the quality of wine based on the compenents of each wine. This paper attempts to predict the quality of red and white wines given alcohol, volatile acidity, sulphates, citric acid, total sulfur dioxide, density, chlorides, fixed acidity, pH, free sulfur dioxide, and residual sugar, using data obtained from the Cortez et al. (2009) case study. Two linear regression models will be computed and analyzed for both red and white wine data sets.

First, upload the data into SAS Studio. For these data I used delimiter as opposed to comma seperated value, and specified the delimiter ';'. Next, begin creating a linear regression model by selecting 'Linear Regression' in the navigation panel of SAS Studio. Specify the data (WORK.RED), the dependent variable (quality), and all eleven continous variables. In the MODEL tab add all variables to the model and select OK. I wanted to include a smooth line in my scatterplots to better identify any patterns. To do this I included the statement PLOTS=RESIDUALPLOT(SMOOTH) in the PRC REG statement, then click run as in Figure 1.

The tables in the RESULTS panel of Figure 1 indicate that 1,599 obervations were read and used, which means there was no missing data. In the Analysis of Variance (ANOVA) table, the F-value calculates if the model is statistically significant, or tests the hypothesis with all independent variable coefficients equal to zero (IBM Corporation, 2022). The p-value is less than 0.0001, which suggests the model is statistically significant. The following table displays root mean squared error (RMSE) and R-squared values. Smaller values of RMSE and higher values of R-squared suggest a better fitted model. The RMSE of 0.64801 and R-squared value of 0.3606 illustrates that the model is not very accurate in predicting the data. In the Parameter Estimates table the t-values were caulated for all independent variables and the model intercept. The t-value calulates the magnitude of variation in the sample data (What are T values and P Values in

Statistics, 2016). Variables with a p-value greater than the significance level (0.05) are considered not statistically significant, and should not be included in the model. Therefore by removing all variables with p-values larger than 0.05, our linear model would be Quality = 21.96521 – 1.08359 * volatile acidtiy – 1.87423 * chlorides + 0.00436 * free sulfur dioxide – 0.00326 * total sulfur dioxide – 0.41365 * pH + 0.91633 * sulphates + 0.27620 * alcohol.

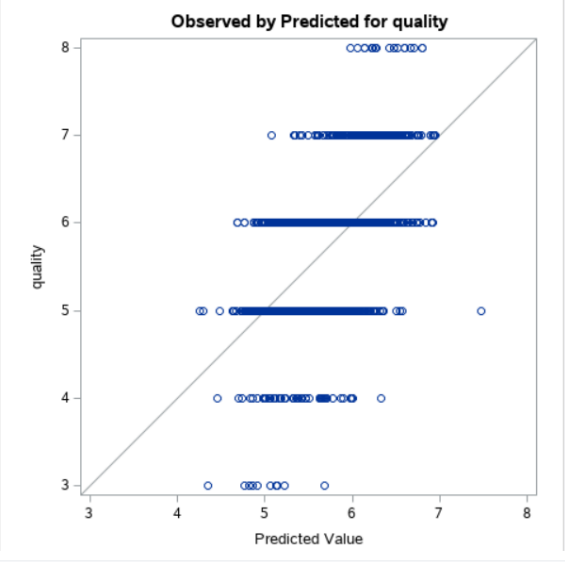
Figure 1: Red Wine Data Set Linear Regression Model with Default Settings



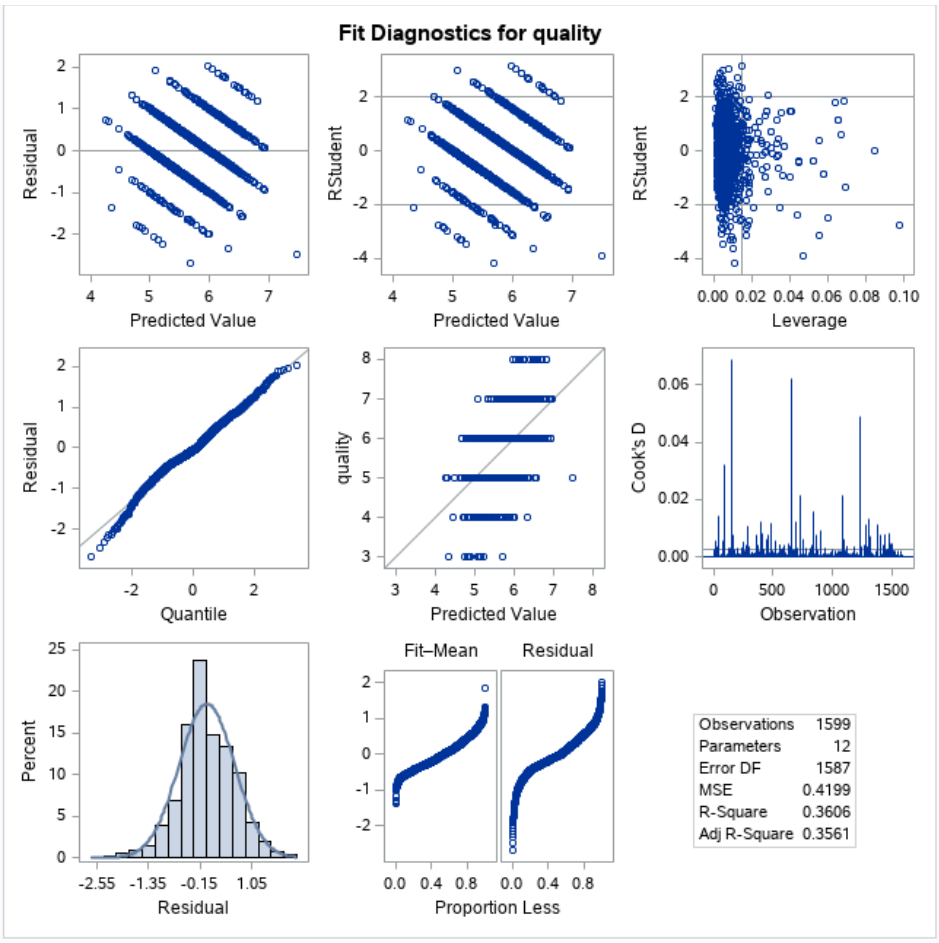
The Observed by Predicted for quality plot in Figure 2 shows that the actual data points do not lie very close to the fitted model, indicating this model is not a very good fit. The Fit Diagnostics graphs illustrate that there may be some potential outliers by the larger values in the Cook's D and Rstudent-Leverage plots. The residuals are approximately normal when viewing the histogram of residuals, the quantile-quantile (Q-Q) plot, and the residual fit-spread plot. Although, the Q-Q plot and the residual fit-spread plot indicate larger values at the extremes.

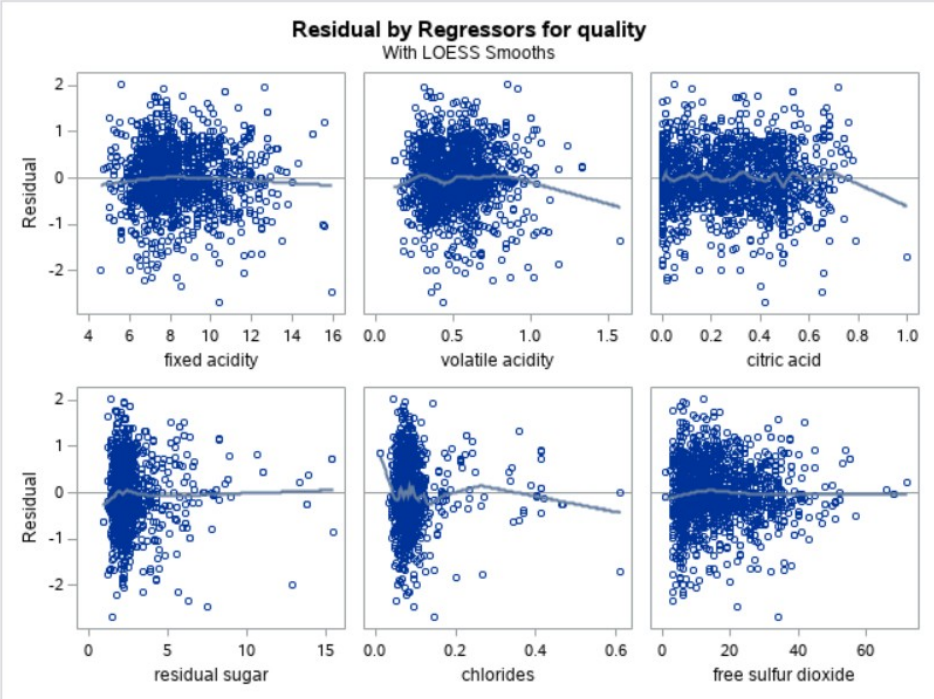
Lastly, it appears that there is a linear trend or pattern in the plot of residual. A pattern in the plot of residual indicates the model is not predicting all of the “signal” in the response variable (Wicklin, 2021, pp.13). A possible solution is adding additional effects to the model or trying a nonlinear model, in an attempt to eliminate any patterns. When viewing the correlation analysis of the independent variables there is a correlation between free sulfur dioxide and total sulfur dioxide, with a moderate correlation of 0.66767 (Goodwin, 2022, p.13). This correlation could be contributing to a potential multicollinearity issue. Lastly, the Residual by Regressors plots for the predictor variables depicts the scatterplots of residuals with LOESS smooths superimposed to observe any potential patterns more clearly. All variables do not appear to have a pattern except pH. It appears that the values are negative at the extremes and approximately zero at the peak. This pattern indicates that the model is not capturing the dependence of the response on pH sufficiently (Wicklin, 2021, pp.24). A possible solution is to try to include an interaction term to model a quadratic response.

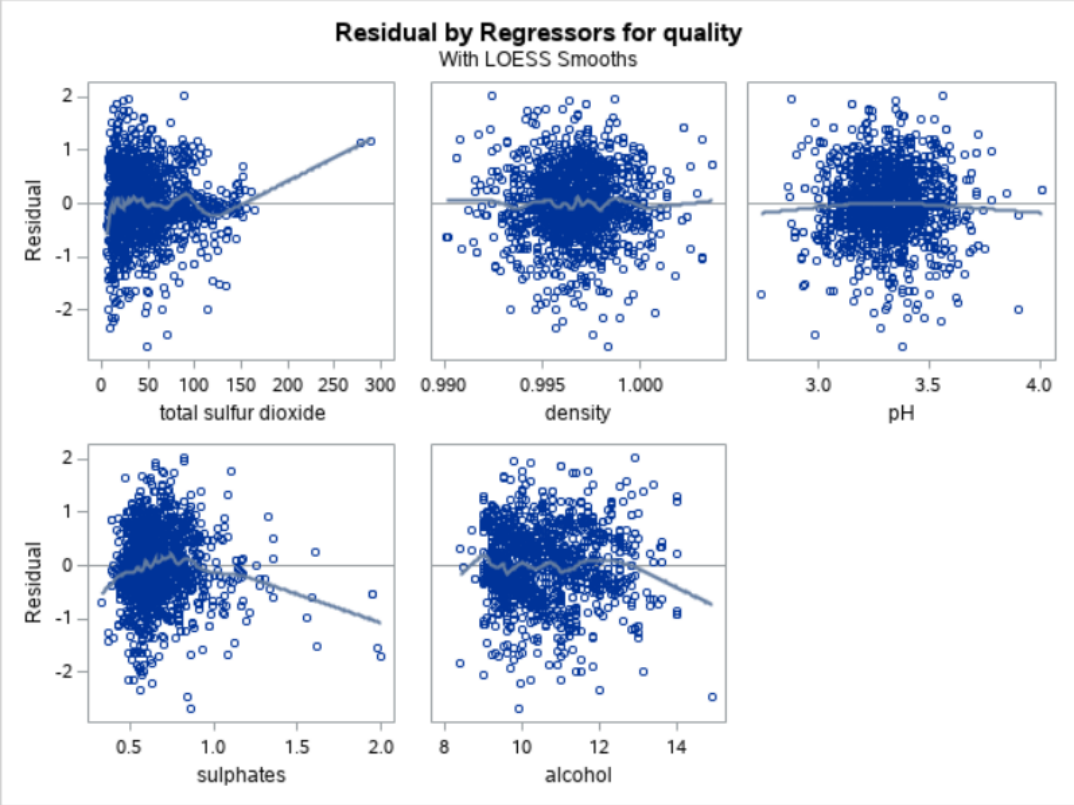
Figure 2: Red Wine Linear Regression Model with Default Settings Output



Module 8: Portfolio Project







The white wine dataset in Figure 3 was performed using the same steps for linear regression as in Figure 2, including adding the LOESS smooths to the residual by regression plots. In the RESULTS tab of Figure 3, the first table illustrates that 4,898 observations were read and used, meaning there are no missing values in this data set. The ANOVA table depicts the calculations for the model's F-value. The p-value of less than 0.0001 shows that this model is statistically significant. The higher RMSE of 0.75136 and the lower R-squared value of 0.2819 demonstrates that this model does not represent the data well. In the parameter estimates table, the variables with p-values larger than the significance level of 0.05 are labelled as not statistically significant and shall not be included in the model. The linear model is represented by Quality = 150.19283 + 0.06552 * fixed acidity – 1.86318 * volatile acidity + 0.008148 * residual sugar + 0.00373 * free sulfur dioxide – 150.28417 * density + 0.68634 * pH + 0.63148 *

sulphates + 0.19348 * alcohol. The large coefficient associated with density represents that changes in density have the largest influence on quality.

The observed by predicted plot shows that the model does not fit the data well because many points to not fall on or near the fitted line. The fit diagnostics show a potential a linear pattern in residual and Rstudent-Leverage plots. There are potential outliers present that require further review, illustrated by the large values shown in Cook's D and Rstudent-Leverage plots. The residuals appear to have an approximately normal distribution illustrated by the histogram, Q-Q plot, and residual fit-spread plot. The Q-Q plot and residual fit-spread plot show higher values at the extremes, indicating this model has larger extremes than a normally distributed data set. Although this model appears to satisfy the normality assumptions, since there is a linear trend in the residual plot and Rstudent-predicted value plots, the data may respond better to a nonlinear model or adding effects. The residual by regressors plots depict random distribution of the error terms for all independent variables. This means the model is accurately representing the dependence of all variables.

Figure 3: White Wine Data Set Linear Regression Model with Default Settings

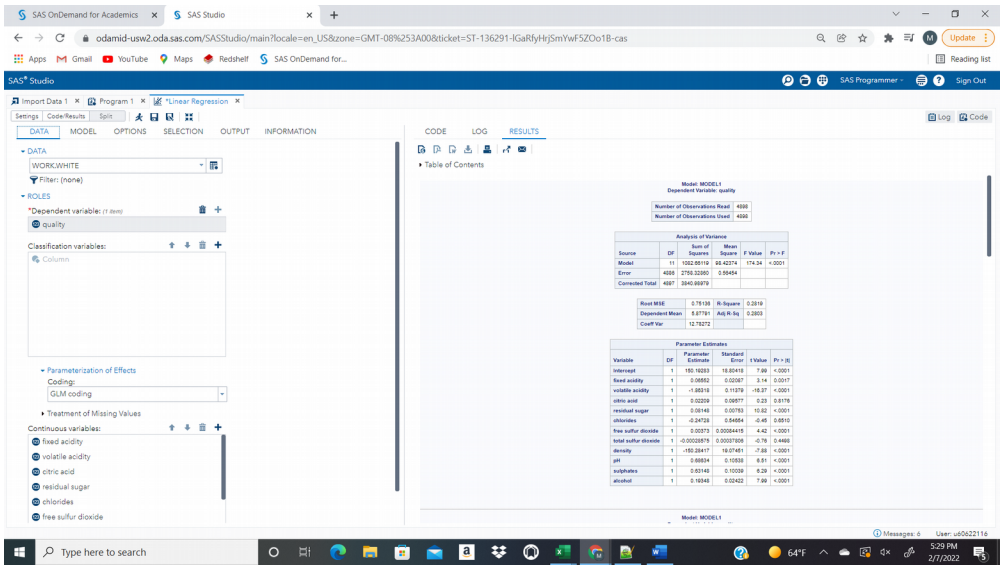
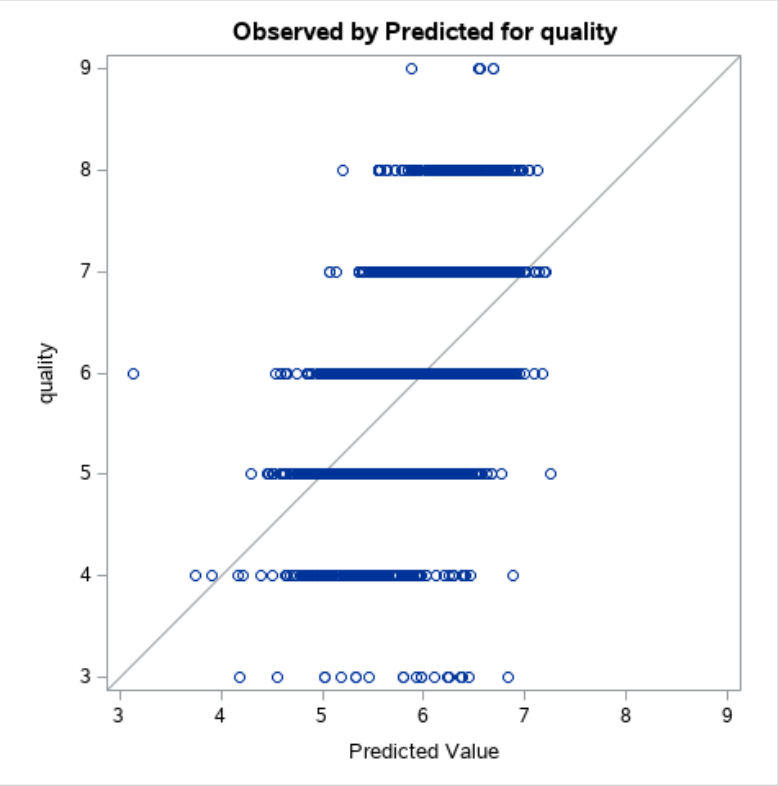
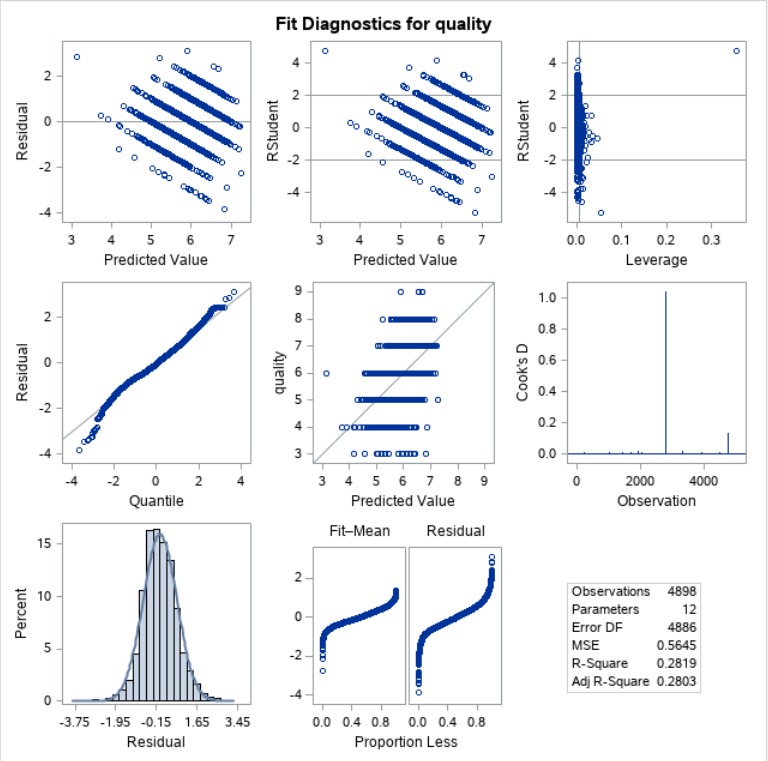
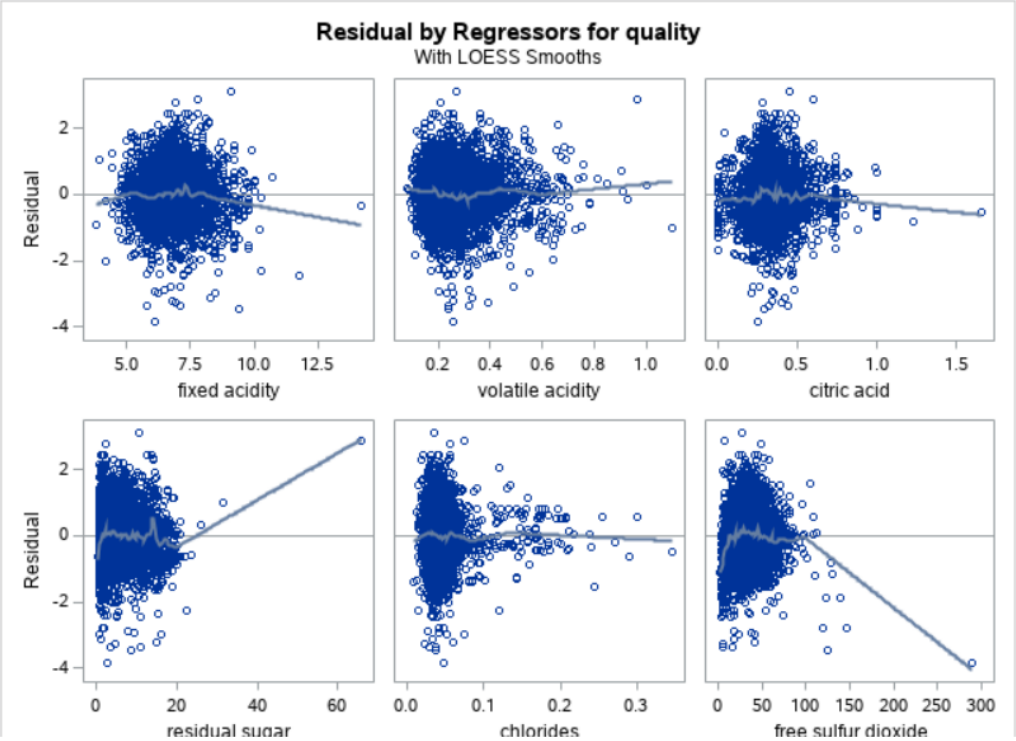


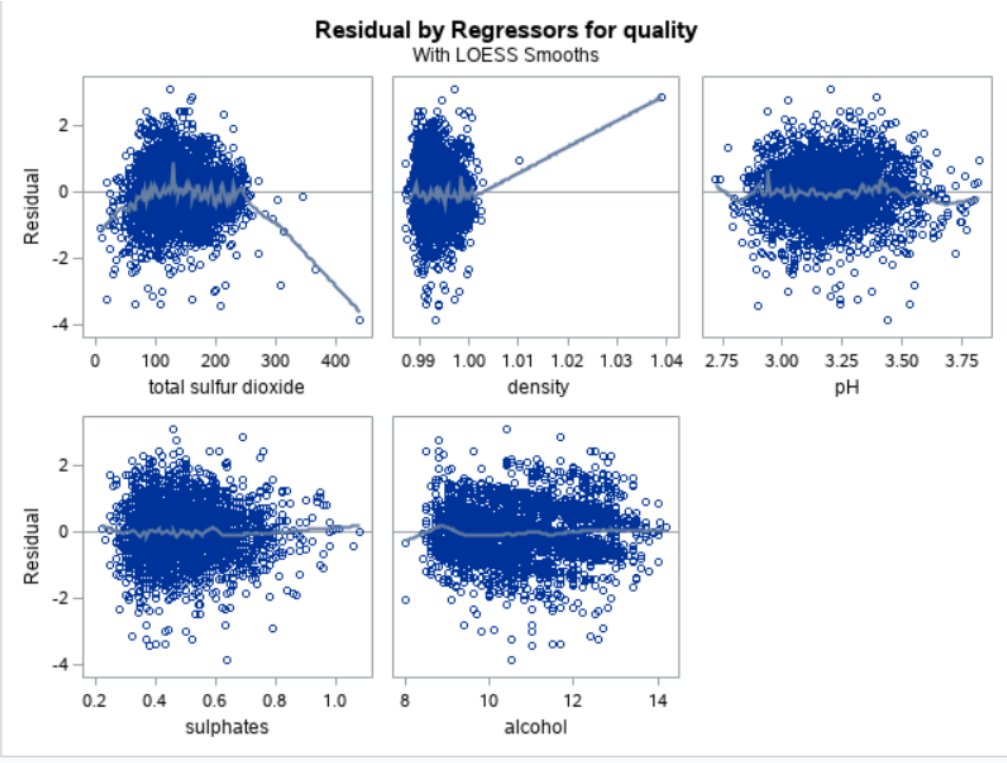
Figure 4: White Wine Linear Regression Model with Default Settings Output





Module 8: Portfolio Project





Forward selection is the process of beginning with just the intercept and adding variables that improve the fit of the model the most (Stepwise Regression, 2016). This procedure attempts to select the variables that will contribute to the best fit. In Figure 4, a linear regression model was created using the forward selection method. To do this in SAS Studio, click Tasks in the navigation panel, select linear regression, select the data (WORK.RED), select the dependent variable (quality), and select all continuous independent variables. In the Model tab, add all independent variables to the linear model. In the Selection tab select Forward selection as the selection method. Additionally, I added the LOESS smooths by selecting code, clicking edit, and typing PLOTS=RESIDUALPLOT(SMOOTH) in the PROC REG statement.

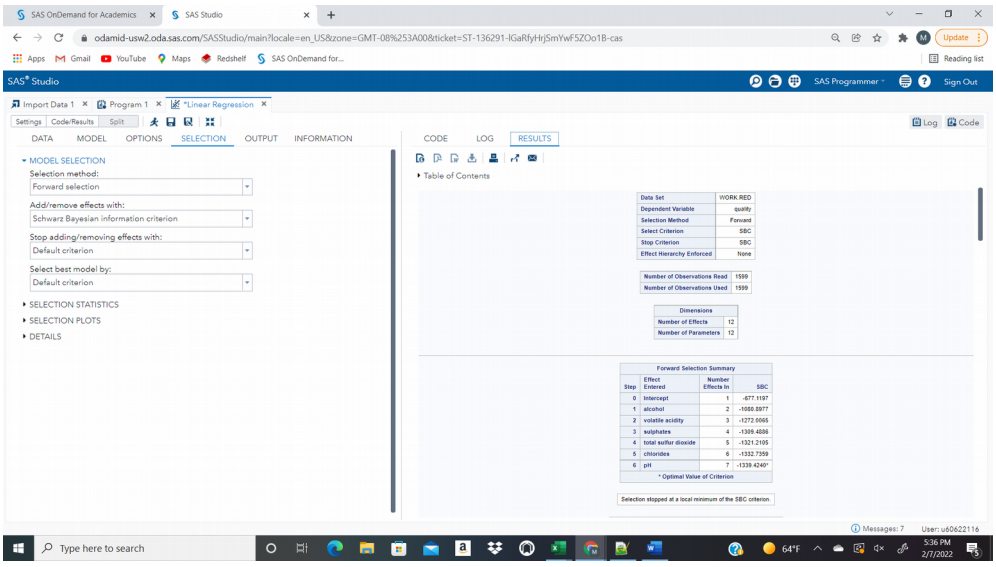
Once the code is ran, the output seen in Figure 4 shows the first table, that describes the data, selection method, and selection criterion. Next, the number of read and used observations were 1,599. The following table shows the number of dimensions, which is twelve for the number of variables. The Forward Selection Summary table indicates the variables added to the model are alcohol, volatile acidity, sulphates, total sulfur dioxide, chlorides, and pH. The algorithm stopped adding variables at free sulfur dioxide as the SBC value would have increased to -1337.7715 by adding the variable.

The Fit Criteria for Quality table indicates the optimal variable selection for the AIC, AICC, SBC and Adjusted R-squared would have been the same six variables, which is not always the case. ANOVA calculations in the following table indicate that the model is statistically significant with a p-value of less than 0.0001, which is smaller than the significance level of 0.05. The RMSE and R-squared values are 0.64870 and 0.3572 respectively. This indicates the model is not doing a good job in predicting the data points. The AIC, AICC, and SBC values that are shown in Figure 4 are values that combine information about SSE, number

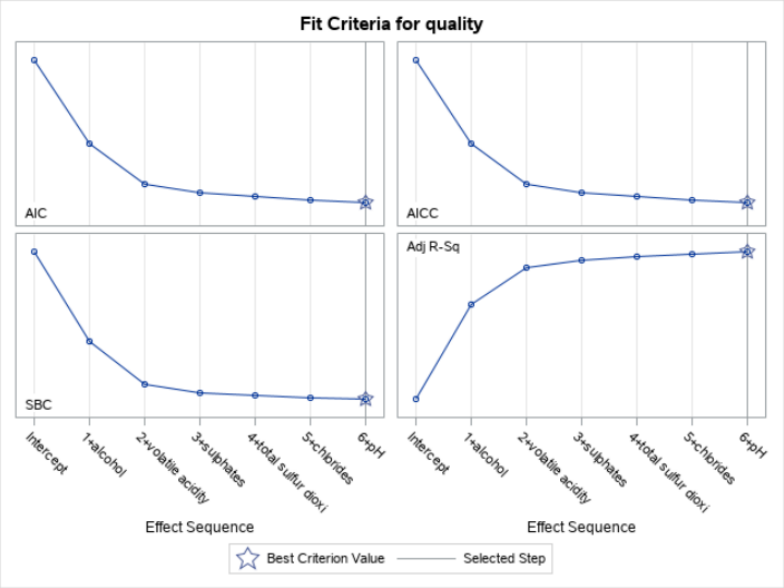
of parameters in the model, and sample size (Information Criteria and PRESS, 2022). They are used to compare the fit of two models, and do not provide any information on their own. Parameter estimates table performs the t-test for each independent variable, and all variables are statistically significant. The model coefficients are also calculated defining the linear model as, quality = 4.295732 – 1.038195 * volatile acidity – 2.002284 * chlorides – 0.002372 * total sulfur dioxide – 0.435183 * pH + 0.888680 * sulphates + 0.290674 * alcohol.

The next plot, Observed by Predicted for quality shows that many of data points do not fall on or near the fitted line, indicating this is a poor model. Fit Diagnostics for quality plot suggest there are more than a few potential outliers, observable by the large data points in the Rstudent-Leverage plot and Cook's D plot. The residuals are approximately normal, illustrated by the Q-Q plot, histogram, and residual fit-spread plot. Residual and Rstudent- predicted value plots still show a linear trend, suggesting additional effects or using a nonlinear model instead. The Residual by Regressor for quality plot does not depict any patterns for the independent variables.

Figure 4: Red Wine Data Set Linear Regression Model with Forward Selection



Stop Details				
Candidate For	Effect	Candidate SBC		Compare SBC
Entry	free sulfur dioxide	-1337.7715	>	-1339.4240



The selected model is the model at the last step (Step 6).

Effects:	Intercept volatile acidity chlorides total sulfur dioxide pH sulphates alcohol

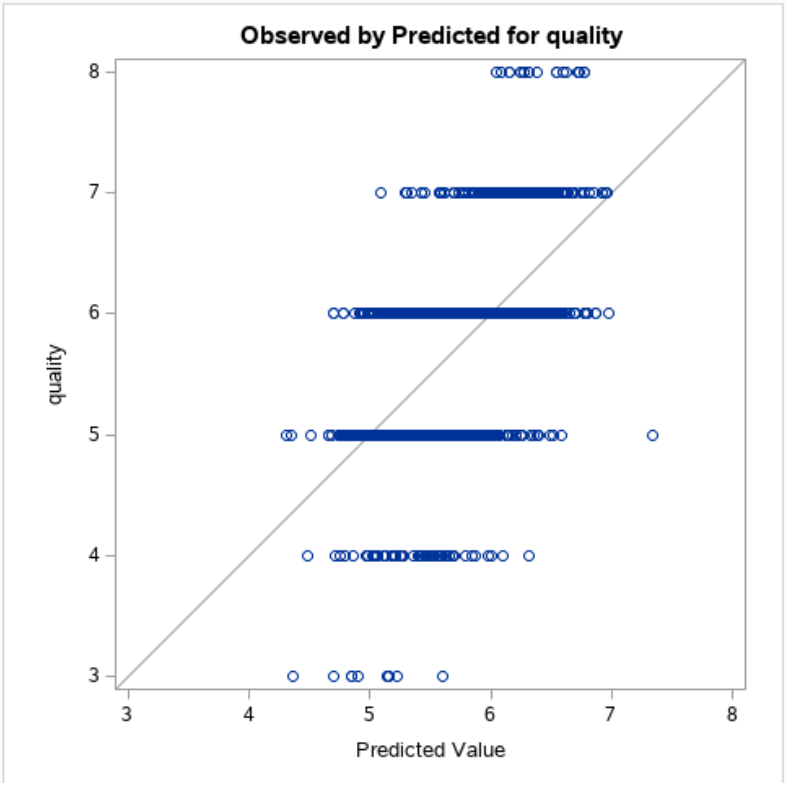
Note: The p-values for parameters and effects are not adjusted for the fact that the terms in the model have been selected and so are generally liberal.

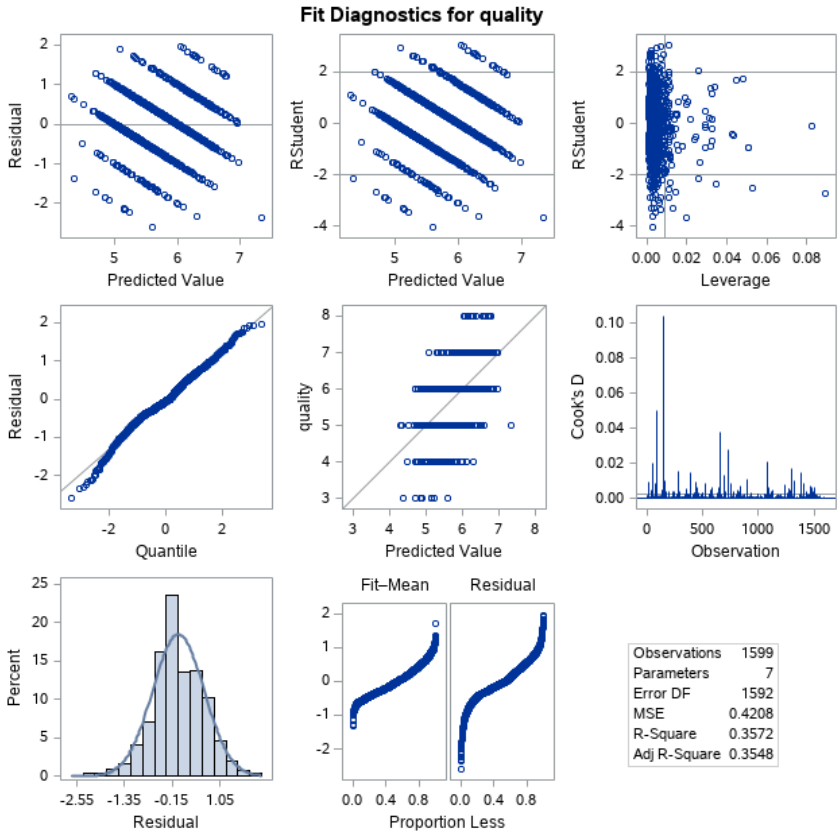
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	372.23391	62.03899	147.43	<.0001
Error	1592	669.93119	0.42081		
Corrected Total	1598	1042.16510			

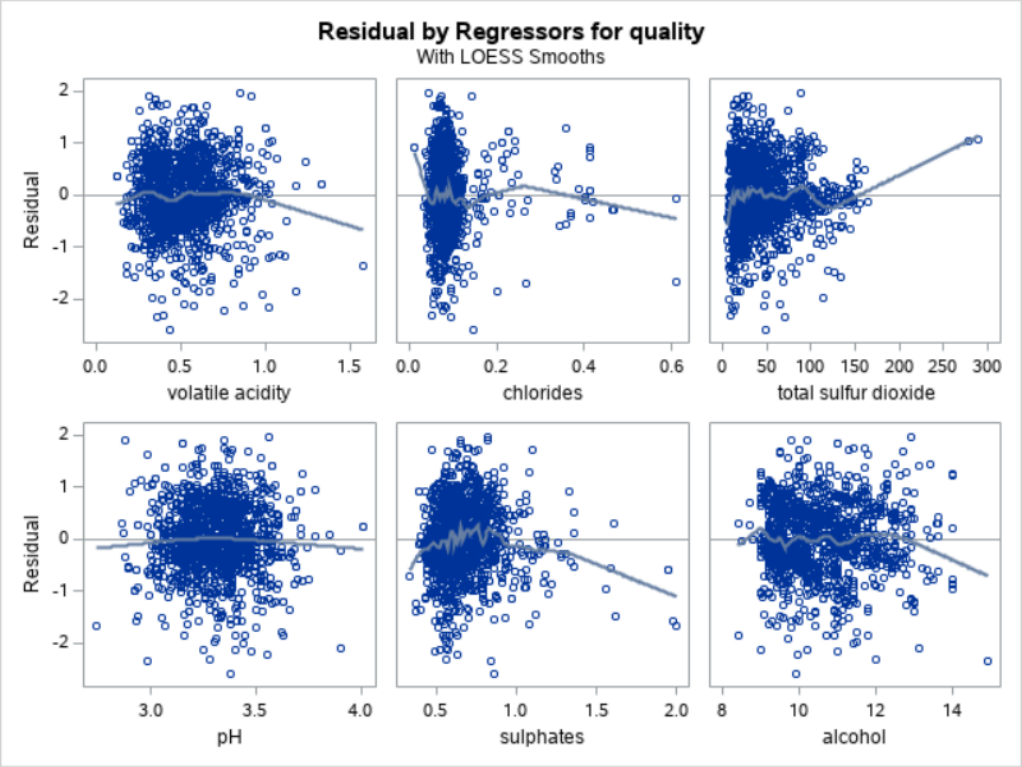
Root MSE	0.64870
Dependent Mean	5.63602
R-Square	0.3572
Adj R-Sq	0.3548
AIC	223.93603
AICC	224.02660
SBC	-1339.42403

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.295732	0.399560	10.75	<.0001
volatile acidity	1	-1.038195	0.100427	-10.34	<.0001
chlorides	1	-2.002284	0.398076	-5.03	<.0001
total sulfur dioxide	1	-0.002372	0.000506	-4.68	<.0001
pH	1	-0.435183	0.116037	-3.75	0.0002
sulphates	1	0.888680	0.110042	8.08	<.0001
alcohol	1	0.290674	0.016811	17.29	<.0001

Module 8: Portfolio Project







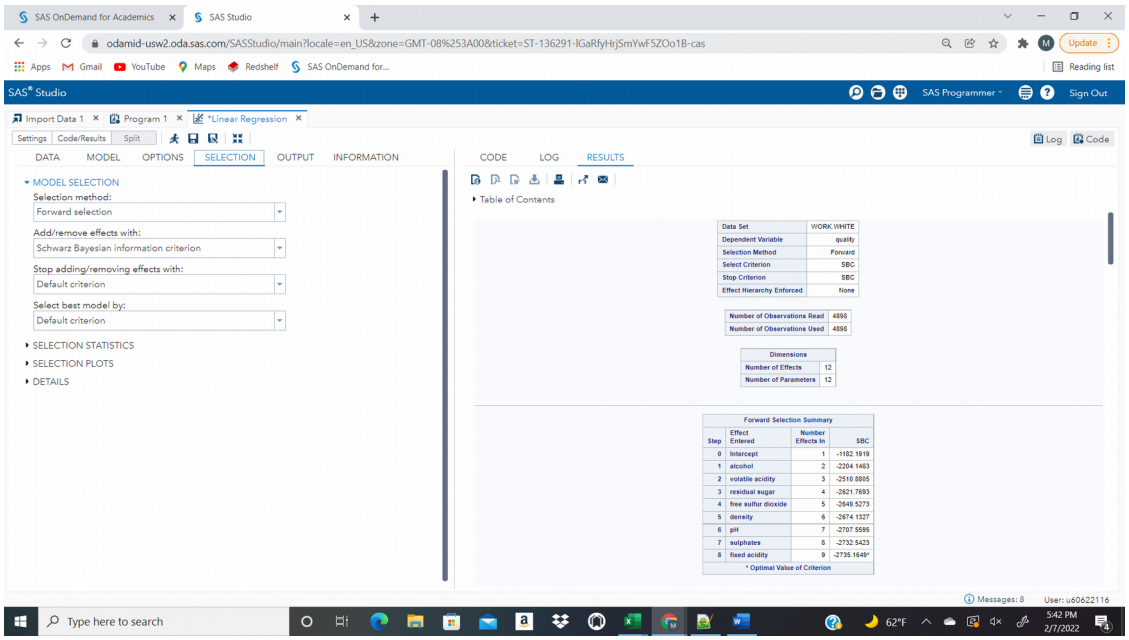
The final model is the white wine forward selection linear model. This model is created using the same steps as in Figure 4 but substituting in the WORK.WHITE data. The results for Figure 5 outline the data used, the selection method, and the selection criterion. This model read and used 4,898 observations with twelve variables. Forward selection deduced that alcohol, volatile acidity, residual sugar, free sulfur dioxide, density, pH, sulphates, and fixed acidity are included in the model. The cumulative SBC values are included in the Forward Selection Summary table. The model stops at total sulfur dioxide because adding this variable would increase the SBC value from -2735.1649 to -2727.2369. The Fit Criteria for Quality plot shows that AIC, AICC, SBC, and adjusted R- squared would select the same eight variables in this model as the optimal criterion value.

ANOVA indicates that the model is statistically significant, meaning the eight independent variables are jointly statistically significant. The RMSE is relatively high at

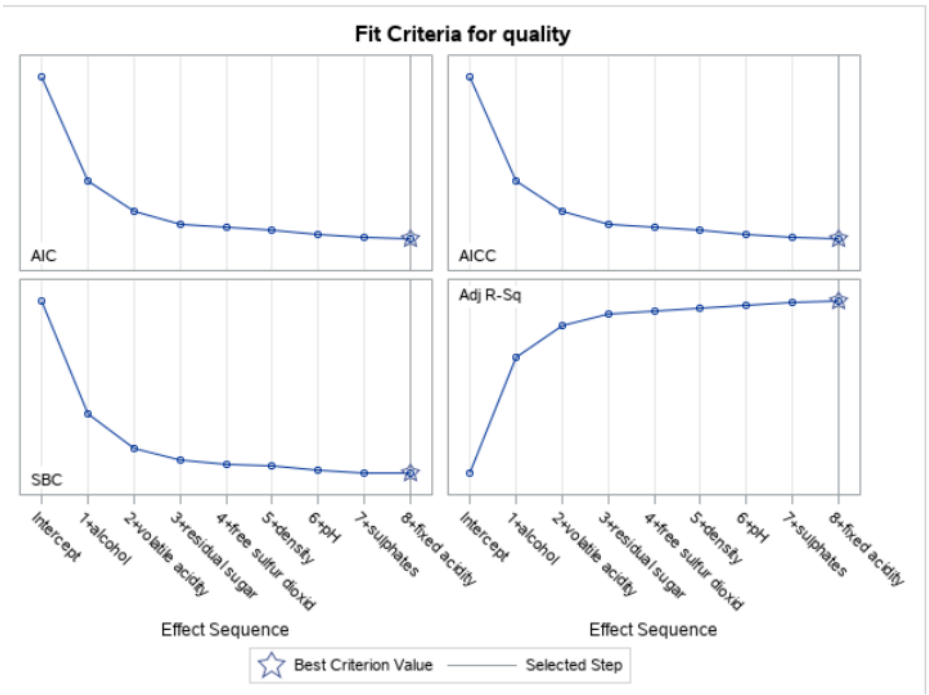
0.75119. R-squared value is 0.2818 which is very low. Both RMSE and R-squared suggest the model is poor. Parameter Estimates table indicates all variables are statistically significant. The model derived from the coefficient estimates is quality = 154.106248 + 0.068104 * fixed acidity – 1.888140 * volatile acidity + 0.082847 * residual sugar + 0.003349 * free sulfur dioxide – 154.291275 * density + 0.0694213 * pH + 0.0628508 * sulphates + 0.193163 * alcohol.

The following plot, Observed by Predicted for quality graph illustrates the poor model fit, as many data points do not lie on or close to the fitted line. There are few outliers suggested by the Cook's D and Rstudent-Leverage plots. The residuals are approximately normal, shown by the Q-Q plot, histogram, and Residual fit-spread plot. There is some deviation at the extremes. As in Figure 2, the residual plot and the Rstudent-predicted plot appear to have a linear trend. Any trend in the residual plot may suggest that model is not capturing everything. Perhaps a nonlinear model or additional effects may result in a better model. Residual by Regressor for quality scatterplots do not show any patterns.

Figure 5: White Wine Linear Regression Model with Forward Selection



Stop Details				
Candidate For	Effect	Candidate SBC		Compare SBC
Entry	total sulfur dioxide	-2727.2369	>	-2735.1649



Selected Model

The selected model is the model at the last step (Step 8).

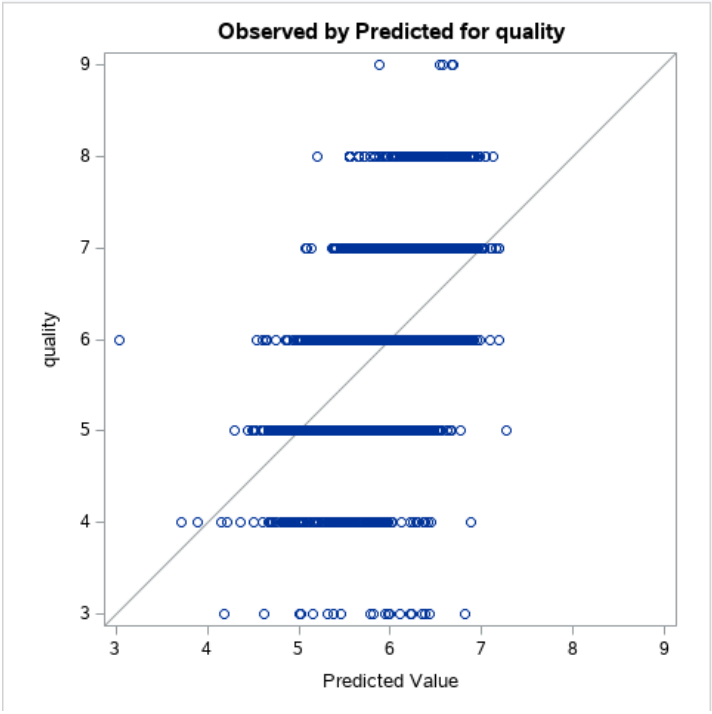
Effects:	Intercept fixed acidity volatile acidity residual sugar free sulfur dioxide density pH sulphates alcohol

Note: The p-values for parameters and effects are not adjusted for the fact that the terms in the model have been selected and so are generally liberal.

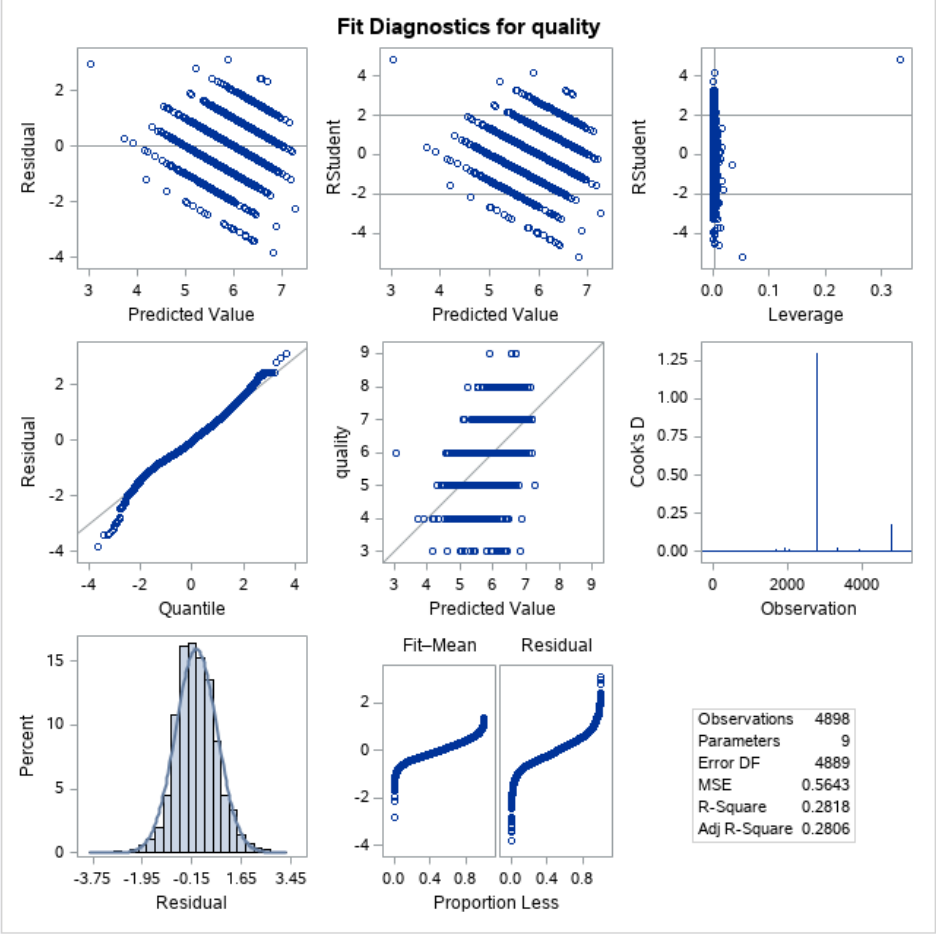
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	1082.20642	135.27580	239.73	<.0001
Error	4889	2758.78338	0.56428		
Corrected Total	4897	3840.98979			

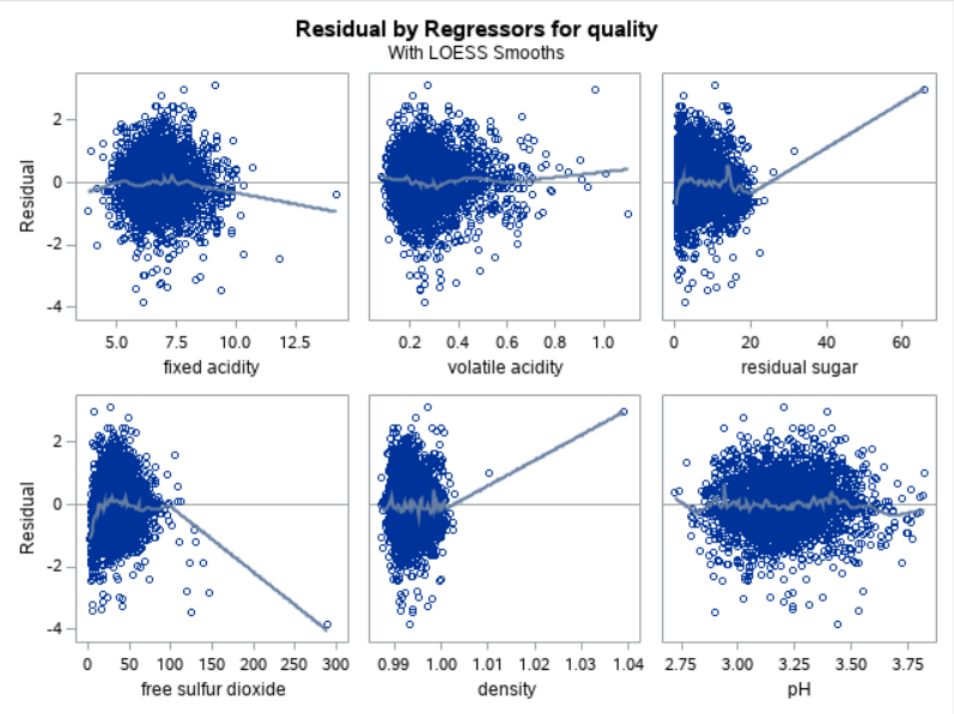
Root MSE	0.75119
Dependent Mean	5.87791
R-Square	0.2818
Adj R-Sq	0.2806
AIC	2106.36588
AICC	2106.41090
SBC	-2735.16488

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	154.106248	18.100130	8.51	<.0001
fixed acidity	1	0.068104	0.020431	3.33	0.0009
volatile acidity	1	-1.888140	0.109509	-17.24	<.0001
residual sugar	1	0.082847	0.007287	11.37	<.0001
free sulfur dioxide	1	0.003349	0.000677	4.95	<.0001
density	1	-154.291275	18.343983	-8.41	<.0001
pH	1	0.694213	0.103351	6.72	<.0001
sulphates	1	0.628508	0.099972	6.29	<.0001
alcohol	1	0.193163	0.024083	8.02	<.0001


Observed by Predicted for quality
quality
Predicted Value
9
8
7
6
5
4
3
3
4
5
6
7
8
9

Module 8: Portfolio Project





One challenge I faced is that I find greater difficulty in reading and interpreting plots that use discrete data. Linear regression models are useful in predicting continuous dependent variables, and discrete numerical dependent variables. Although, linear regression may not be the best model for every data set. The white and red wine data sets illustrate that the linear regression model performed poorly, despite using forward selection. Through the analyses conducted in this paper we are able to identify potential solutions for creating the best model for this data set, such as constructing a nonlinear model. For future analyses with these data I would potentially try a regression tree analysis and compare the performance of those models with the linear models.

References

IBM Corporation. (2022). *F value*. IBM. Retrieved February 8, 2022, from https://www.ibm.com/docs/en/cognos-analytics/11.1.0?topic=terms-f-value

Information Criteria and PRESS. (2022). Penn State Eberly College of Science. Retrieved February 11, 2022, from https://online.stat.psu.edu/stat501/book/export/html/971

Goodwin M., (2022, January 20). Module 5: Portfolio Milestone 2.

NCSS. (2016). *Stepwise Regression*. Retrieved February 11, 2022, from https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Stepwise_Regression.pdf

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

What Are T Values and P Values in Statistics? (2016, November 4). Minitab. Retrieved February 8, 2022, from https://blog.minitab.com/en/statistics-and-quality-data-analysis/what-are-t-values-and-p-values-in-statistics#:%7E:text=The%20t%2Dvalue%20measures%20the,evidence%20against%20the%20null%20hypothesis.

Wicklin, R. (2021, March 24). *An overview of regression diagnostic plots in SAS*. SAS Blogs. Retrieved February 8, 2022, from https://blogs.sas.com/content/iml/2021/03/24/regression-diagnostic-plots-sas.html