Christian Rosse

11/5/2024

DSC 323

**Wine Quality Analysis**

This report focused on a dataset consisting of 1599 observations of the red wine variants of a Portuguese wine company by the name of Vinho Verde. The dataset consists of 12 total variables ([D21](#Bookmark28)) . These variables include acidity levels such as volatile acidity and fixed acidity. According to WineEnthusiast.com volatile acidity is a measure of the wine’s gaseous acids in which the higher the acidity the higher likelihood of spoilage, and fixed acidity are the acids that preserve the wine’s natural color ([29](#Bookmark29)). There are sulfur measurements in the dataset that include free sulfur and total sulfur. Free sulfur dioxide is a compound that protects the wine from oxidation, according to an article posted by Aaron Essary of Oklahoma State University, and total sulfur includes the free sulfur dioxide and any other sulfuric gas compounds ([L1](#Bookmark23)). The dataset includes flavor indicators such as residual sugars and citric acid. The residual sugar indicates the amount of sugar left in the wine after fermentation and citric acid is added to the wine to increase acidity or complement flavor, according to an article published by UC Davis. Other variables within the dataset include pH level, chlorides, sulphates, and density which describe the wine’s chemical composition ([L2](#Bookmark24)). The final variable is the quality of wine. This quality variable is quantified by sensory data and scaled on a 1-10 rating system. A Vinho Verde red wine with a 10 quality rating indicates a wine with the best quality and a wine with a 1 quality rating indicates the worst quality.

This data model will be testing the dependency of the quality variable compared to the other variables given within the dataset. The null hypothesis for this experiment should be:

*There is no correlation between the wine’s quality and any of the 11*  *independent variables within the dataset.*

Correspondingly, the alternative hypothesis that this experiment will test reads:

*There is, in fact, a correlation between wine quality and one or more of the*  *independent variables within the dataset.*

The purpose of this examination is to potentially build a model that will help companies such as Vinho Verde make data driven decisions about the chemical composition of their wine in order to maximize the quality of their product.

Observing the 1599 observations, it is assessed that there are no missing values and because of that there is no need to sift out any empty or null observations before data exploration begins. The goal will be to produce a linear regression model as the dependent variable is numeric. There will be no need to create dummy variables on the data since all the variables within the dataset are numeric values. First, a histogram ([D1](#Bookmark1)) will be created in order to examine the distribution of the wine’s quality throughout the dataset. The results showed a unimodal, normal distribution but with a significant proportion of the wine qualities either a 5/10 or a 6/10. The 5-number summary ([D3](#Bookmark3)) shows that the interquartile range is from 5-6 so the values in the histogram outside of the peak are deemed as potential outliers and extreme points. Because of the condensed peak, the distribution of wine quality is deemed normal. The mean of all the qualities is 5.636 and the median is 6 ([D3](#Bookmark3)). The minimum quality observed is 3 while the maximum quality observed is 8 ([D3](#Bookmark3)).

With no significant need to run a transformation or create dummy variables, linear regression will be used to generate the full model to predict wine quality. Running a goodness of fit test on the full model shows a high F-value of 81.35 and the model’s overall p-value is statistically significant ([D4](#Bookmark4)). This indicates that the null hypothesis can be rejected as at least one of the predictors explain trends in the wine quality. Meanwhile, the R-Squared is 0.3606 ([D4)](#Bookmark4). The R-Squared value indicates that 36.06% of the wine quality’s variance is explained by all 11 variables, and this indicates those 11 variables within the full model hold moderately low explaining power, as the model stands. On top of that, it is seen that variables fixed acidity (0.0689), density (0.1473), residual sugar (0.0756) and citric acid (0.1284) were statistically insignificant as they have p-values > 0.05 ([D4)](#Bookmark4). Before splitting the model into train and test sets, checking for outliers and multicollinearity is needed.

With a large group of predictor variables such as the 11 in the dataset, the potential for multicollinearity remains a factor. It is deemed that running a correlation matrix and getting the full model’s VIF values should be sufficient in catching multicollinearity. Observing the correlation matrix ([D2](#Bookmark2)), it is seen that fixed acidity shows linear trends with density, pH and citric acid. The highest VIF values show moderate collinearity with values of 7.8 for fixed acidity and 6.4 for density ([D4](#Bookmark4)). Due to the VIF values not being above 10 to indicate extreme collinearity, it is deemed to exercise caution of the variables and make note but wait for the results of future model selection.

The last step this experiment will exercise in the data exploration stage will be finding, and dealing with, outliers and influential points. Generating a Cook’s D vs Studentized residual ([D22](#Bookmark30)) chart in SAS will reveal three data observations that shows both high influence and are considered outliers. Due to the data set being fairly large, there are little worries that removing 3 points will affect the predicting power of the model, so it is deemed that those influential points shall be removed. Removing those points improves the R-Squared value by 1.25%, going from 0.3606 ([D4](#Bookmark4)) to 0.3731 ([D5](#Bookmark5)). As a rule of thumb, if a round of outlier removal yields a 2% or less R-Squared improvement, outlier removal shall be halted. So, with outliers removed, multicollinearity noted and checked, and goodness of fit confirmed, the model can now be split into training and testing sets to develop the best possible model for this data.

Because of the large size of the dataset, it is decided that an 80/20 split is appropriate for this model. Doing so yields 1277 observations for the training set, and 319 unseen observations for model testing. It is decided that there should be two different types of model selection in order to keep options for potential models open. The first model will be selected using the stepwise method. The reasoning for deciding the stepwise method is because it combines aspects of both forward and backward selection methods while operating in a step by step fashion. The confidence in stepwise is high because it completes all of the necessary trials to sort out the best predictor variables to use in the final model. The second selection method is the Adjusted R-Squared method. The reason for choosing this method for model number two is that it gives the analyst a laundry list of potential model options ranked by the highest R-Squared value. This gives more freedom to the analyst to choose which model may be the best to train with.

Model #1 was selected stepwise in 7 steps ([D7](#Bookmark8)). This yielded a potential final model consisting of 7 out of the 11 original predictor variables. All predictor variables in this model were statistically significant, with p-values falling under the 0.05 threshold ([D8](#Bookmark9)). On top of that, the overall model p-value was statistically significant as well, confirming the rejection of the null hypothesis. The R-Squared (0.3683) and Adjusted R-Squared (0.3648) ([D8](#Bookmark9)) dropped slightly lower than the full model, and the RMSE (0.64077) ([D8](#Bookmark9)) grew slightly higher. Analyzing the residual plots ([D9](#Bookmark10)), there are good indicators of constant variance and independence throughout most of the predictor plots. However, free sulfur dioxide and total sulfur dioxide’s residual plots ([D10](#Bookmark11)) show a slight “V” shape indicating potential issues with variance and independence. The normality plot ([D11](#Bookmark12)) follows a straight-line pattern for the most part, indicating that the residuals are normally distributed.

As Model #2 was decided using Adjusted R-Squared model selection, there were a few options available to choose from. The highest Adjusted R-Squared yielding model included all 11 variables ([D23](#Bookmark31)). It was decided to not choose this one as it is previously known that 4 out of the 11 variables are insignificant predictors. Next, the differences in Adjusted R-Squared between the second highest (0.3653) and third highest (0.3652) was that the second highest model had one predictor extra and a 0.01% increase in Adjusted R-Squared ([D23](#Bookmark31)). It is recognized that the fewer predictor variables possible, the better, so the third highest model was opted for. This model had 8 predictor variables and a slightly lower R-Squared value of 0.3652 ([D12](#Bookmark13)) compared to the full model’s 0.3731 ([D5](#Bookmark6)). For Model #2, 36.52% of the variance is described by the 8 predictor variables compared to the 37.31% of the full model’s variance that is described by all 11 predictor variables. As this model has 8 predictor variables compared to model #1’s 7, the only difference between the two models is the addition of citric acid as a predictor. The residual plot of the added citric acid ([D13](#Bookmark14)) shows proper variance and independence due to the randomness of the points scattered within the residual plot. All of the other residual and normality plots between the two models remain the same so all of the previous analysis of Model #1’s residuals can be carried over to here as well. Looking at the p values of the variables within Model #2 ([D12](#Bookmark13)), they added citric acid has a p value of 0.1798 which happens to be slightly insignificant as it is above the 0.05 threshold. This will be an important bullet point to consider once choosing which final model to use on predictions.

Transitioning into the model testing phase, for Model #1, the testing results show an RMSE (Root Mean Standard Error) value of 0.6368 ([D16](#Bookmark17)). This indicates that on average, the model’s predictions are about 0.6368 units away from the actual values. This is a slightly lower error value than the Model #1’s training phase’s 0.64077 RMSE value([D8](#Bookmark9)). This suggests that the testing on this particular set was slightly more accurate in predicting wine quality using the model. This model’s testing MAE (Mean Absolute Error) value is 0.49445 ([D16](#Bookmark17)). This MAE suggests that the average prediction is about 0.49445 units away from the true value. The MAE is a lower error value than the RMSE due to the fact that RMSE penalizes more for outliers as it squares the value before averaging. As MAE is just purely the average absolute value of the distance between predicted and actual, it gives the analyst a better picture of how accurately the model is truly predicting onto the test set. The Pearson Correlation Coefficient matrix displayed from our model testing shows that relationship between the wine’s quality and y-hat is 0.61428 ([D16](#Bookmark17)). This number indicates that the relationship between the wine’s true quality and predicted quality is moderately correlated in a positive direction. Furthermore, when the wine’s true quality goes up it is likely the predicted quality will go up as well. This correlation metric can be used to find the testing models R-Squared. R-Squared is found by squaring the 0.61428 to get an R-Squared value of 0.3773. This value says that 37.73% of the testing model’s variance can be described by the model’s 7 predictor values of volatile acid, chlorides, free sulfur dioxide, total sulfur dioxide, pH, sulphates and alcohol. The Adjusted R-Squared value is found for this model by adding the R-Squared value to the formula and getting an Adjusted R-Square value of 0.3746. This number suggests that 37.46% of the testing model’s variance is described by the seven significant predictors, with a deduction to penalize potential overfitting. Lastly, determining the cross validated R-Squared by subtracting the testing model’s R-Squared with the testing model’s R-Squared (0.3773 - 0.3683), we get a difference of 0.009. This shows that the Model #1’s testing phase described slightly more variance within the model’s testing set compared to the training set. As it is around the same value as the original R-Squared value from the model’s training phase, it can be determined that there is no overfitting nor underfitting present and the testing model is performing how it should be upon unseen data.

Analyzing Model #2, the testing results show an RMSE (Root Mean Squared Error) value of 0.63815 ([D17](#Bookmark19)). This indicates that on average, the model’s predictions are about 0.63815 units away from the actual values. This model, also, has a slightly lower error value than the Model #2’s training phase’s 0.64057 RMSE value ([D12](#Bookmark13)). Just like Model #1 testing, this suggests that the testing on this particular set was slightly more accurate in predicting wine quality using the model. This model’s testing MAE value is 0.49551 ([D17](#Bookmark19)). This MAE suggests that the average prediction is about 0.49551 units away from the true value. The Pearson Correlation Coefficient matrix displayed from our model testing shows that relationship between the wine’s quality and y-hat is 0.61216 ([D17](#Bookmark19)). This number indicates that the relationship between the wine’s true quality and predicted quality is moderately correlated in a positive direction. Furthermore, when the wine’s true quality goes up it is likely the predicted quality will go up as well. This correlation metric can be used to find the testing models R-Squared. R-Squared is found by squaring the 0.61216 ([D17](#Bookmark19)) to get an R-Squared value of 0.37474. This value says that 37.47% of the testing model’s variance can be described by the same 7 predictor values as Model #1, but with the addition of citric acid as a predictor variable. The Adjusted R-Squared value is found for this model by adding the R-Squared value to the formula and getting an Adjusted R-Square value of 0.37159. This number suggests that 37.16% of the testing model’s variance is described by the seven significant predictors, with a deduction to penalize potential overfitting. Finally, finding the cross validated R-Squared by subtracting the testing model’s R-Squared with the testing model’s R-Squared (0.37474 - 0.369), we get a difference of 0.00564. This shows that the Model #2’s testing phase described slightly more variance within the model’s testing set compared to the training set. As it remains close to the original R-Squared model from the training phase, it can be determined that there is no overfitting nor underfitting present and the testing model is performing how it should be upon unseen data.

Once all the metrics have been accumulated, the best model needs to be determined. Comparing the training models, Model #2 has a higher R-Squared (0.3691) ([D12](#Bookmark13)) than Model #1’s R-Squared (0.3683) ([D8](#Bookmark9)), and Model #2 has a higher Adjusted R-Squared (0.3652) ([D12](#Bookmark13)) than Model #1’s Adjusted R-Squared (0.3648) ([D8](#Bookmark9)). This could suggest that on a larger dataset, such as training, the use of the citric acid predictor could be useful in describing variance within the model. Model #2’s training RMSE (0.64057) ([D12](#Bookmark13)) is marginally lower than Model #1’s (0.647) ([D8](#Bookmark9)). This could also indicate that with a larger dataset, the citric acid predictor could potentially be useful for reducing the average distance between predicted versus actual within the model. So, this means that on strictly the training set, Model #2 performs ever so slightly better even with the addition of the slightly insignificant predictor in citric acid. However, Model #1 does, in fact, perform better on the testing dataset as it records a lower RMSE (0.63685) ([D16](#Bookmark17)) compared to Model #2’s (0.63815) ([D17](#Bookmark19)), and a lower MAE (0.49445) ([D16](#Bookmark17)) compared to Model #2’s (0.49551) ([D17](#Bookmark19)). This indicates that not including citric acid within the model, reduces the average distance between the predicted values and the actual values. This is in line with the notion that citric acid in itself is an insignificant predictor and could provide noise and bias to the results. On a similar note, the R-Squared for Model #1’s testing phase (0.3773) ([D16](#Bookmark17)) and Adjusted R-Squared (0.3746) ([D16](#Bookmark17)) are both higher than Model #2’s R-Squared of (0.3747) ([D17](#Bookmark19)) and Adjusted R-Squared (0.3716) ([D17](#Bookmark19)). Comparing the cross validated R-Squared values, Model #1 has improved the R-Squared by slightly more than Model #2, though both models saw slight improvement in R-Squared when tested on the smaller unseen dataset.

Due to both models being very marginally different in certain areas, there are a few key factors that will sway the decision. Though both models saw improvement when predicting on unseen data, Model #1’s lack of the citric acid predictor variable proved to be slightly more effective when predicting on unseen data. This, combined with the fact that citric acid in itself is not a significant predictor, can sway the higher amount of confidence in Model #1 going forward. Even with Model #2’s better performance on the training phase, the discrepancy between the two is low enough to where it can be chalked up as a draw. And while same notion can be implied toward the small discrepancies between the two testing performances, Model #1’s fewer variables can be the edge needed in making this decision. Fewer variables lead to a more simple and more generalized model, and Model #1 is what is needed to better predict red wine quality.

Model #1 will be used to represent the final model equation in order to determine wine quality based on the seven predictor variables. The intercept (4.43408) ([D8](#Bookmark9)) represents the base value of the wine’s quality if all the other variables are equal to zero. If the amount of each of the predictor variables are equal to zero, theoretically, our model will determine that the wine’s quality is 4.4/10. Though, as chemical variables such as chlorides and alcohol are necessary for a beverage to be defined as wine, there is no anticipation that this such case will occur.

Volatile acid (-1.01498) ([D8](#Bookmark9)) represents that for every unit of volatile acidity that the wine has, the quality will decrease by –1.01498. As discussed in the introduction, the higher amount of these acids increases the risk of spoilage within the wine, so the inverse relationship between volatile acidity and quality mirrors that trend.

Chlorides (-1.97471) ([D8](#Bookmark9)) represents that for every unit of chloride that is increased within the wine, the quality rating will go down almost 2 ticks. Chlorides themselves are an indicator of salty flavor within the wine’s taste. Furthermore, this trend represents that the saltier the wine begins to taste, the worse the quality is perceived.

Free sulfur dioxide (0.00546) ([D8](#Bookmark9)) and total sulfur dioxide (- 0.00353) ([D8](#Bookmark9)) represent interesting relationships. Free sulfur shows that for every unit of free sulfur dioxide added, the quality increases by 0.00546. However, with every increase unit of any sulfur dioxide molecule, the quality decreases by 0.00353. This implies that the wine’s quality is the best, with a higher ratio of free sulfur dioxide to the total amount of sulfur dioxide. According to Aromadictionary.com ([L4](#Bookmark26)), high amounts of sulfur dioxide can make the wine taste metallic-like and bitter. This is represented well in the model equation, as free sulfur has a necessary task in reducing risk of oxidation, but the more sulfur there is, the more bitter it tastes, the wine’s quality benefits the most from a fine balance between the two. This is why these two may not be as collinear as originally anticipated due to the important nature each variable, by itself, holds.

For pH (-0.51056) ([D8](#Bookmark9)), the model equation says that for every unit pH is increased, the quality of wine goes down half a tick. On average, the pH for wine is fairly low around 3. On the pH scale, 3 represents a fairly acidic taste which lines up with the wine’s inherent acidic nature. Although, as pH increases, like the sulfur dioxide, the wine begins to become more basic and more bitter. The trends suggest that more acidic wine is more flavorful and considered better quality than if it was to become bitter tasting.

Sulphates (0.88339) ([D8](#Bookmark9)) represents the highest positive factor in the equation for increasing the wine’s quality. If the sulphate level is increased by 1 unit, the wine’s quality rating would increase by almost a full point. Sulphates level is indicative of the wine’s shelf life as they are crucial in preserving the wine’s flavor. The term “ages like fine wine” refers to the role sulphates play in keeping the wine’s color and taste fresh. This makes sense as to why appropriate levels of sulphates would be indicative of the wine’s increase in quality.

The last variable in the equation is alcohol (0.29757) ([D8](#Bookmark9)). As the model’s equation suggests, ever one-unit increase in the wine’s alcohol level, the wine’s quality will grow by about 0.3. According to daily.sevenfifty.com ([L5](#Bookmark27)), higher alcohol levels can round out the wine’s flavor and mesh all of the flavors together cohesively. This mirrors the model’s positive relationship between the wine and the quality.

All in all, the variables with the highest positive effect on the quality of the wine were sulphates and alcohol. The most negatively influential predictor variables included in the final model were chlorides and volatile acidity.

Using the final model, predictions can be made on a sample wine. For this assessment, the wine was created with 0.63 volatile acidity, 0.08 chlorides, 19 free sulfur dioxide, 33 total sulfur dioxide, 3.67 pH, 1.19 sulphates, and 9.7% ABV (alcohol by volume) ([D18](#Bookmark20)). Running the final model on this sample observation, the model predicts the wine to be 5.6878/10 quality ([D19](#Bookmark21)). The mean of the dataset is 5.636 ([D3](#Bookmark3)), so the wine predicted on is ever so slightly above average. The 95% confidence level mean interval on this prediction has a lower bound of 5.5284 and an upper bound of 5.8471 ([D20](#Bookmark21)). This suggests that with 95% confidence, the average wine quality of a wine with this characteristics falls between 5.5284 and 5.8471. So, either slightly above or below the mean of the dataset. The 95% confidence level prediction interval has a lower bound of 4.4234 and an upper bound of 6.9522 ([D20](#Bookmark21)). This means that with 95% confidence a wine with similar characteristics falls between a quality rating of 4.4234 and 6.9522. Using this model, the feedback it could give to the winery that is developing this particular wine could be to decrease the pH slightly, increase the alcohol level and try to even out the ratio of the sulphates in order to achieve a higher quality wine potentially.

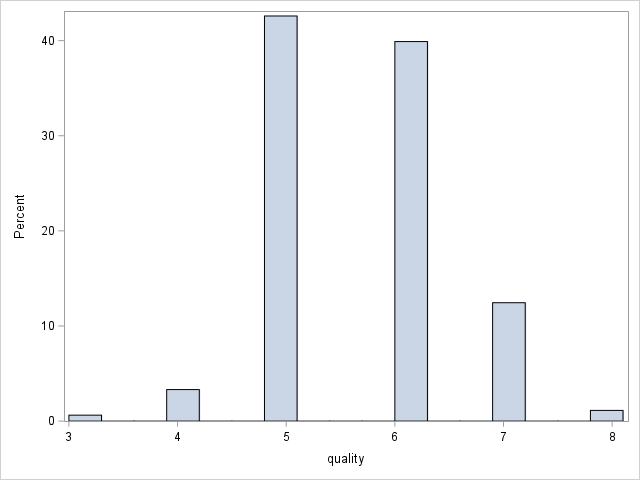
In conclusion, the model that was found was determined using accurate techniques to assess the data and the trends within. However, the model that was found would not be considered a good model for predicting wine quality. Reasons to believe that include a moderately low R-squared value. The final R-Squared of 0.3683 indicated that less than 40% of the wine quality’s variance happened to be described by the predictor variables meaning that more than 60% is unexplained. This could be troubling when trying to convince someone to rely on this model to predict wine accurately. On a similar note, the prediction intervals had a width of about 2.5 on average. This interval spreads 25% of the entire 1-10 scale and since the minimum of wine quality observed within the dataset was 3 and the maximum quality observed within the dataset was 8 ([D3](#Bookmark3)), this width spans 50% of the scale of wine observations trained on. Due to these reasons, this particular model should not be relied on for precise prediction for the wine’s quality as the relationship between the predictor variables and the quality don’t happen to explain one another sufficiently.

With that being said, tests such as this can be important for understanding how elements within the wine react with one another and form a final product. It gives the analyst a scope on how to potentially go about improving a wine’s quality and what steps to possibly consider when doing so.

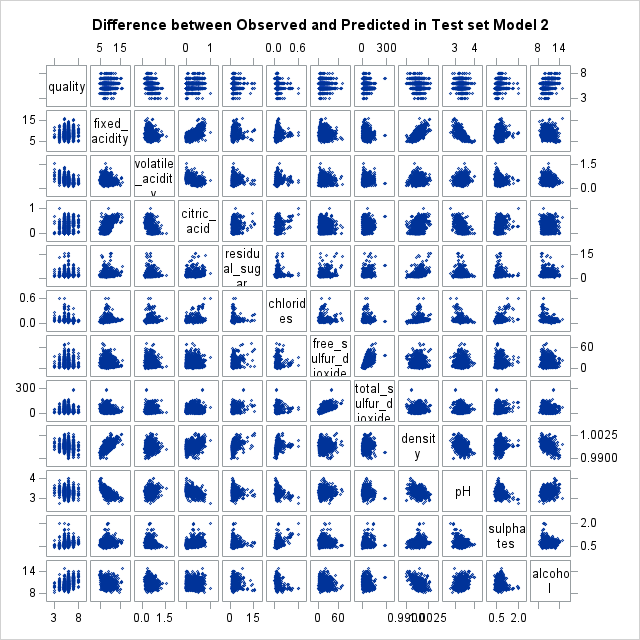
Potential ways in which this exploration can be furthered could be to experiment with logistic regression in order to examine how different variables influence the probability that a wine is good quality versus bad quality. Another test that can be explored is to group together the free sulfur dioxide and total sulfur dioxide into one ratio value that represents what percentage of the sulfur dioxide within the wine is considered free. Lastly, the training testing splits could be played with to see if the increased performance was simply luck on the smaller sample, or would it be successful on a slightly larger sample such as a 75/25 split.

**APPENDIX:**

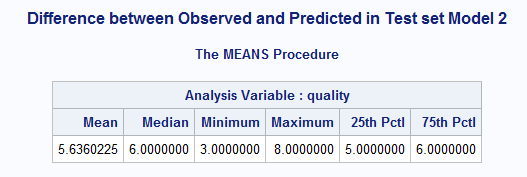
D1



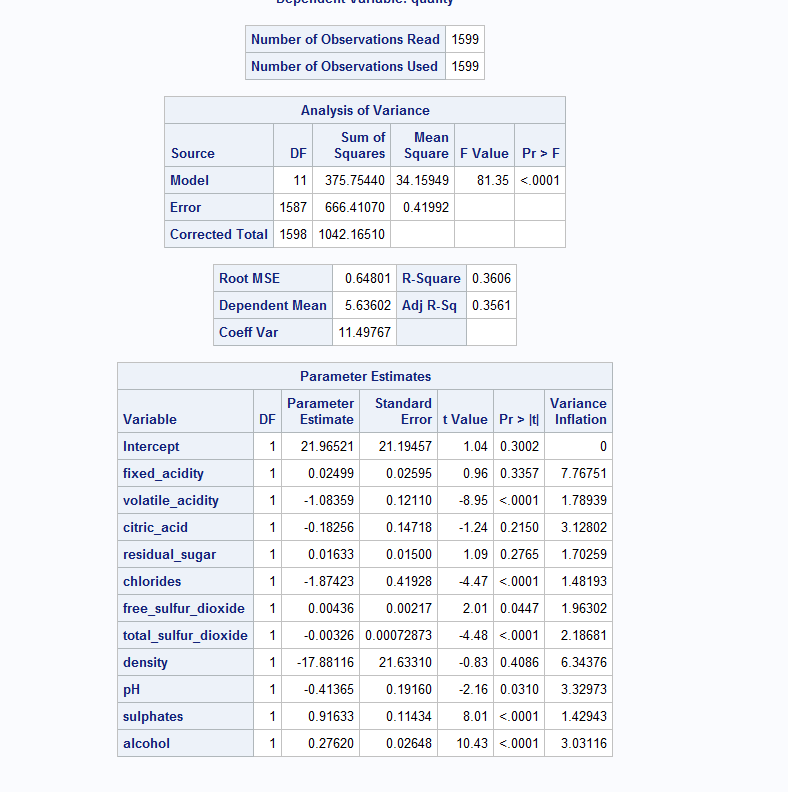
D2



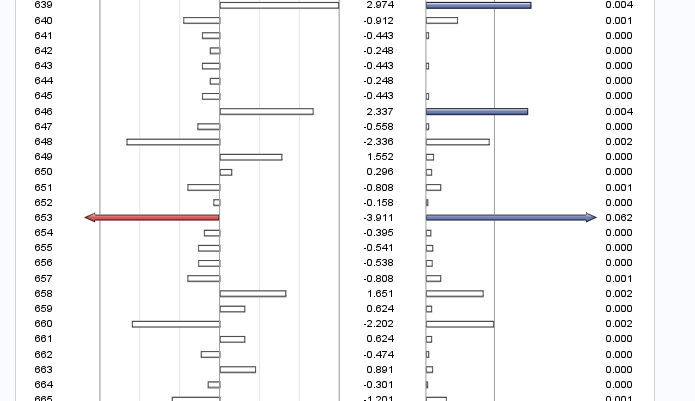
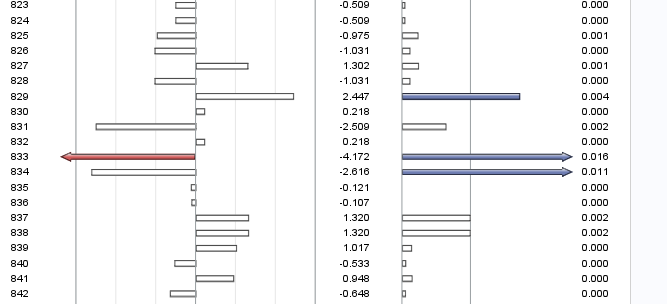
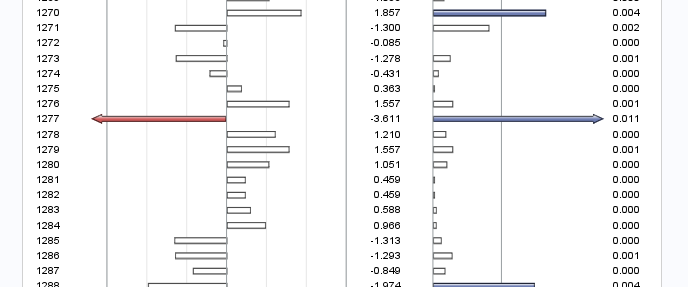
D3



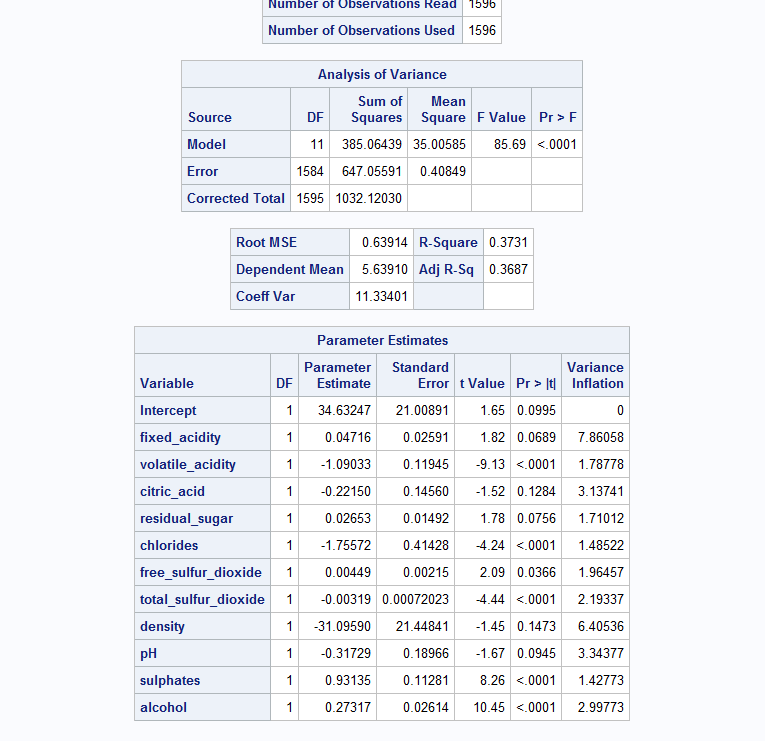
D4



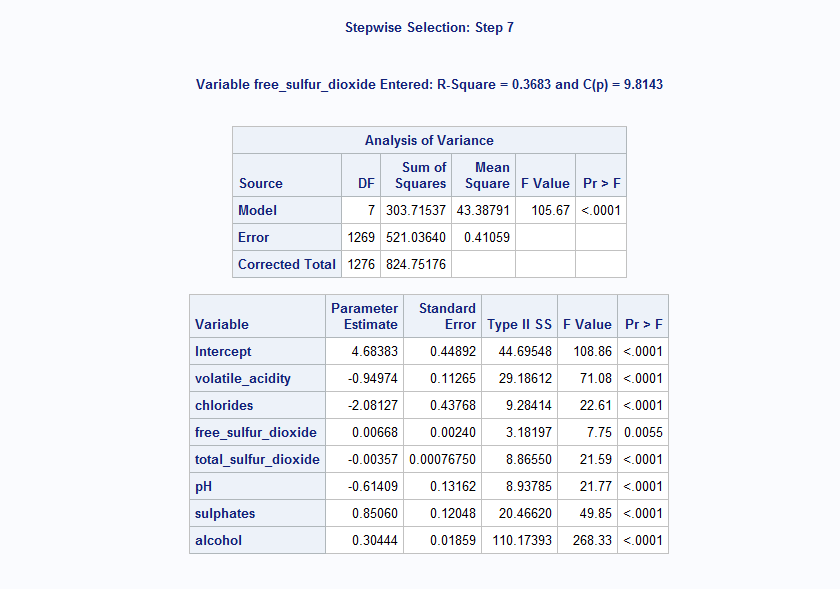
D22

D5



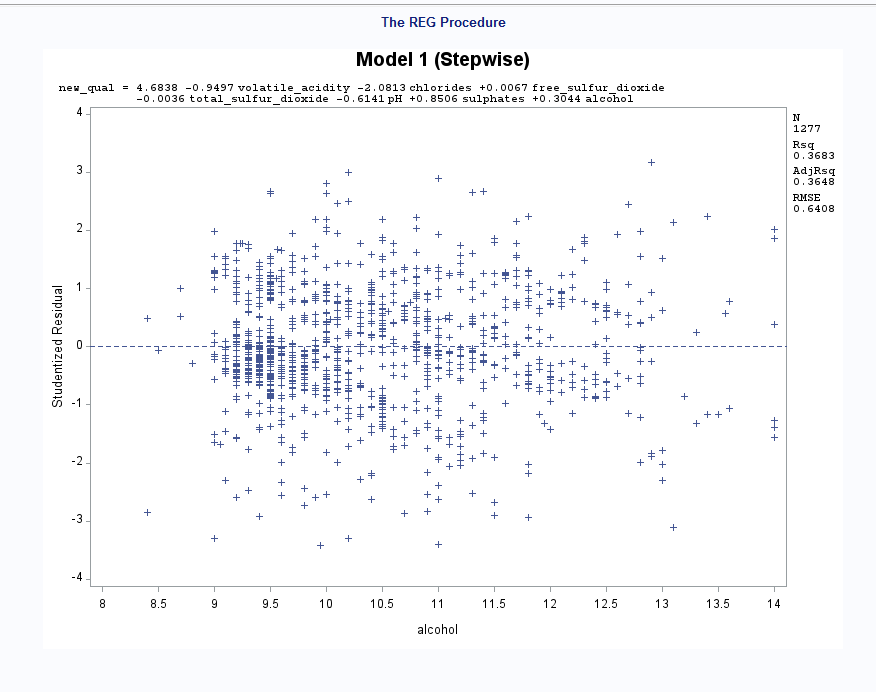
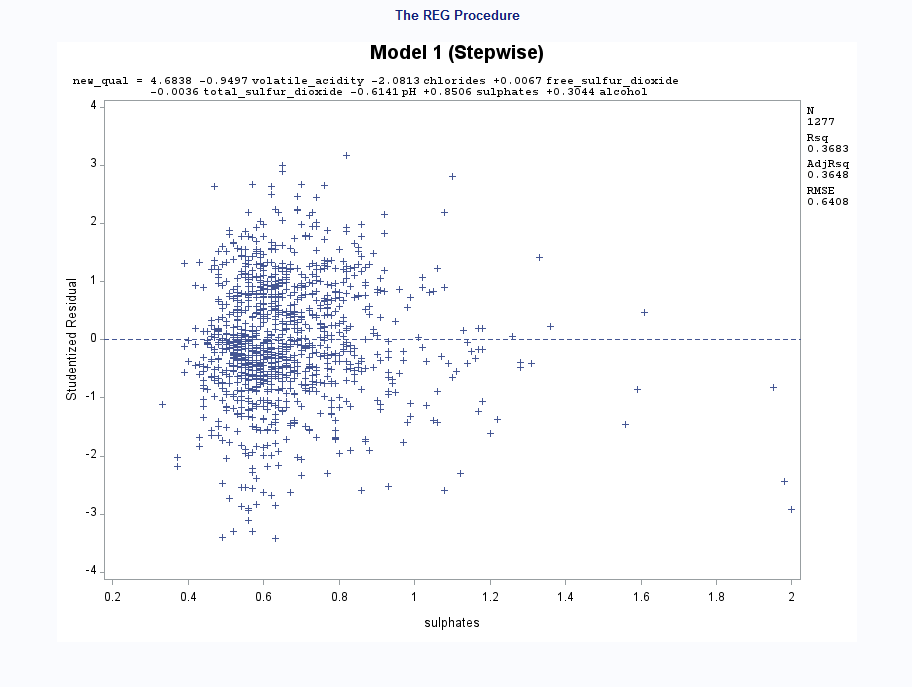
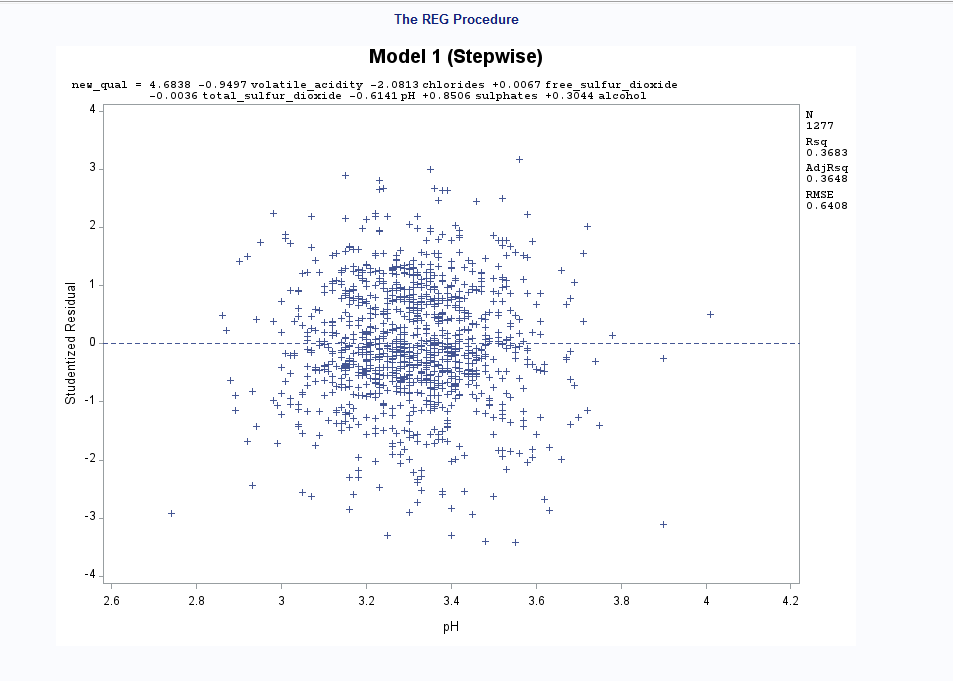
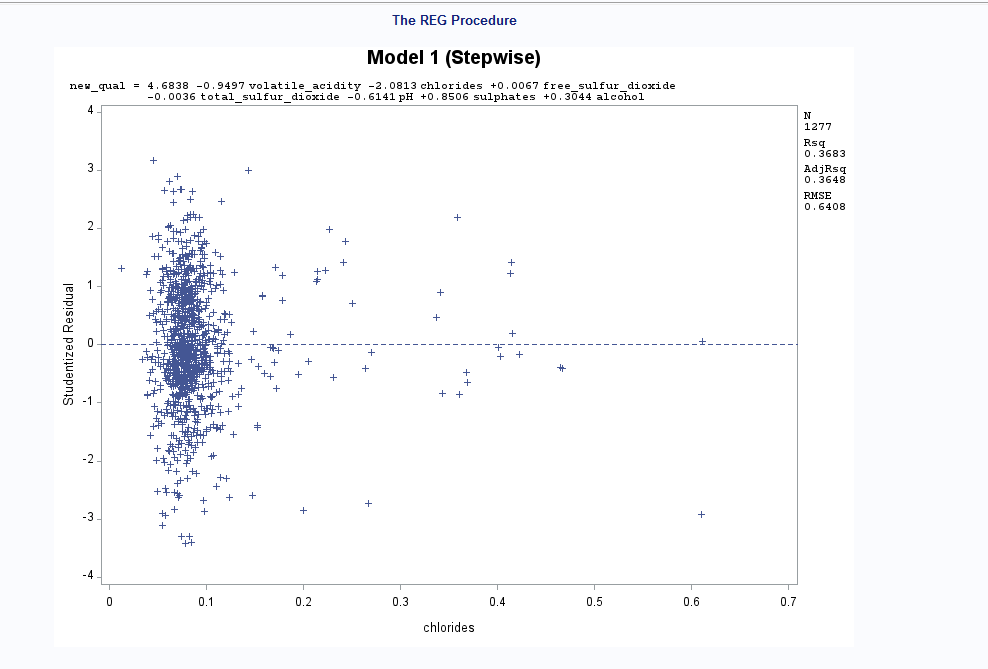
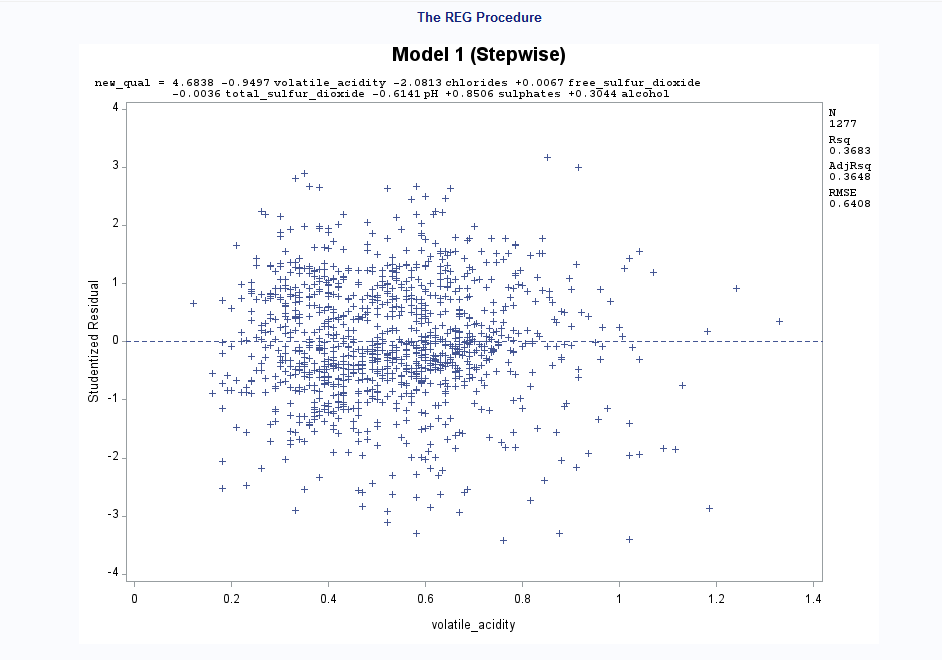
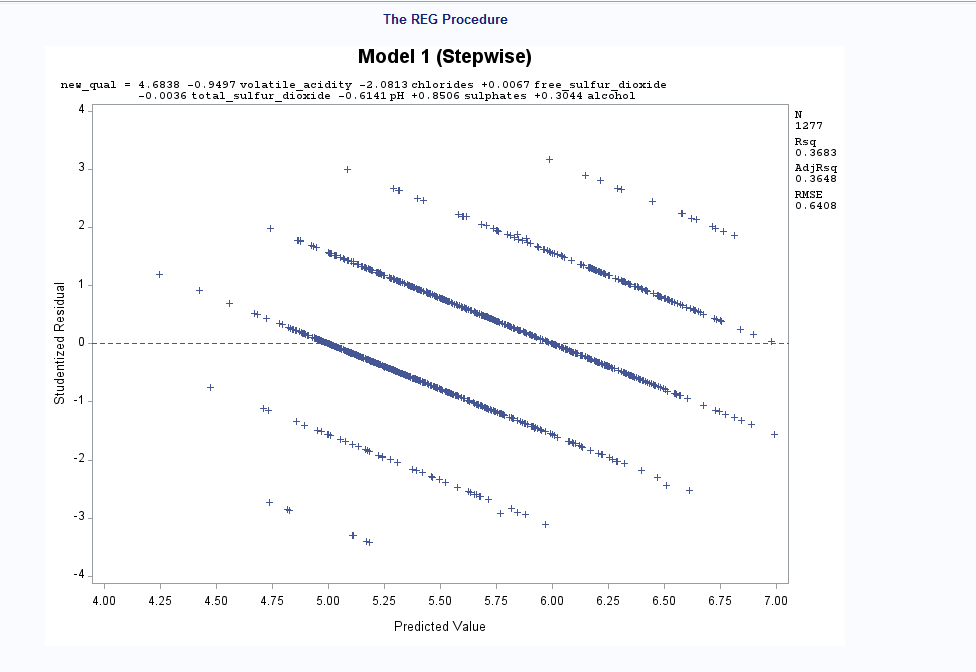
D6



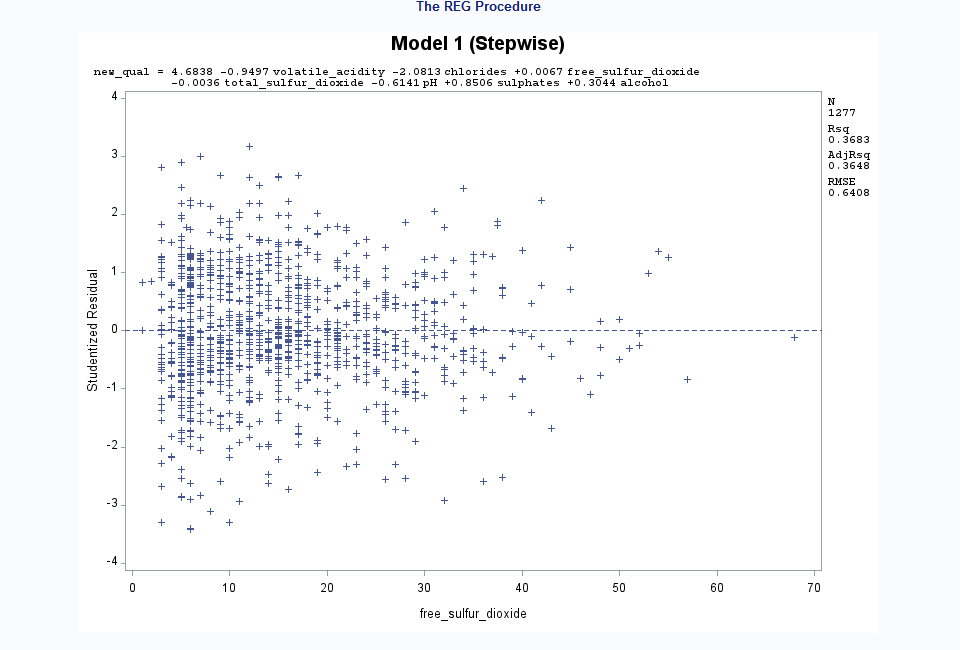
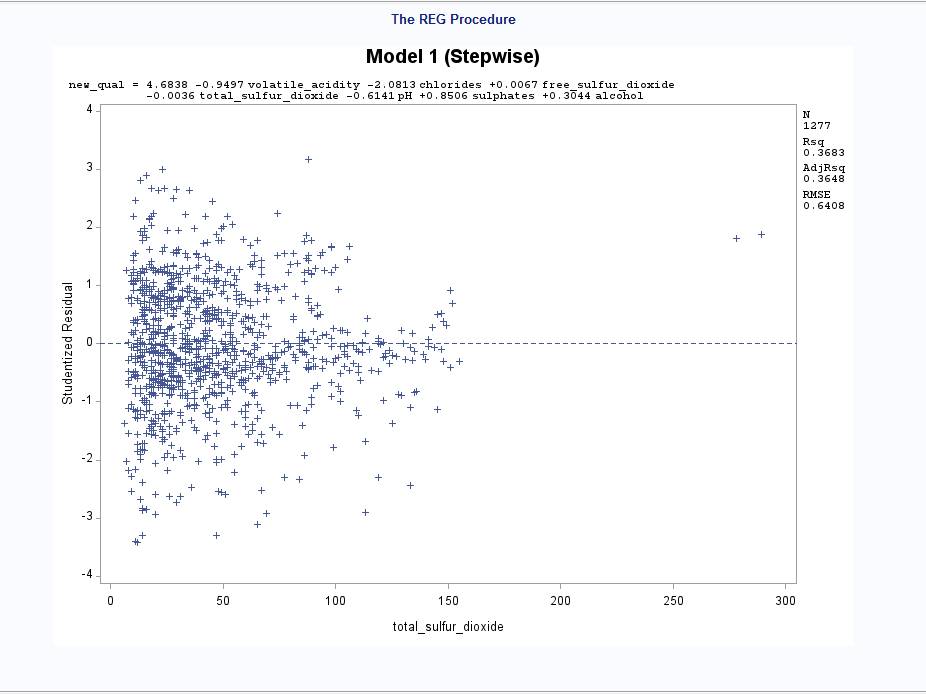
D7



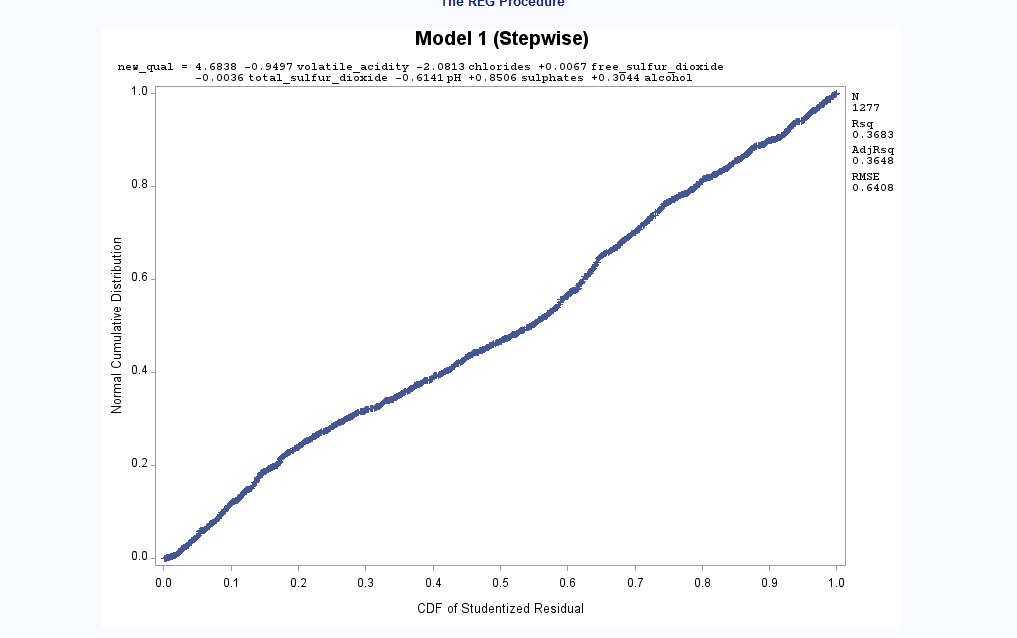
D8



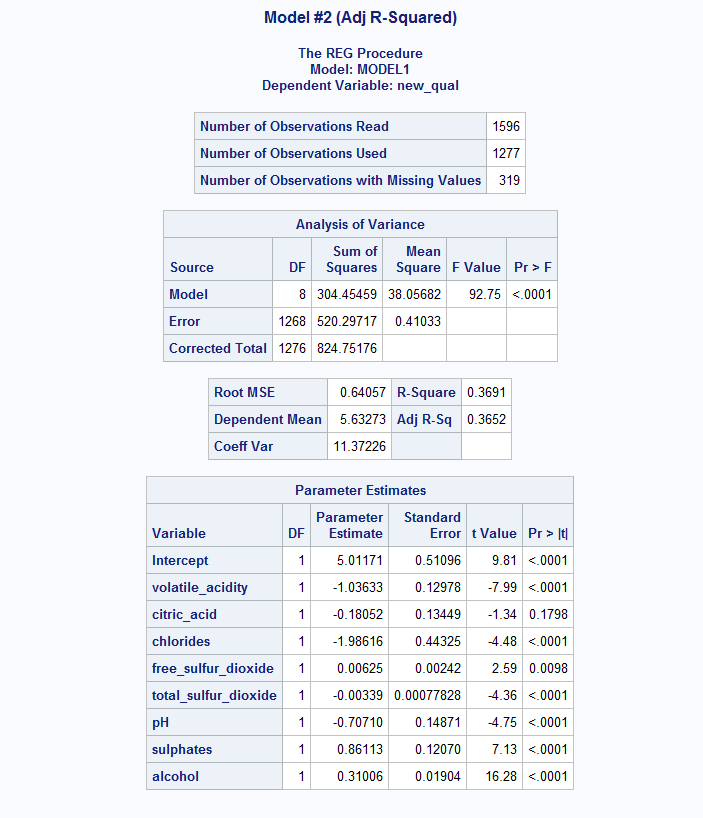
D9



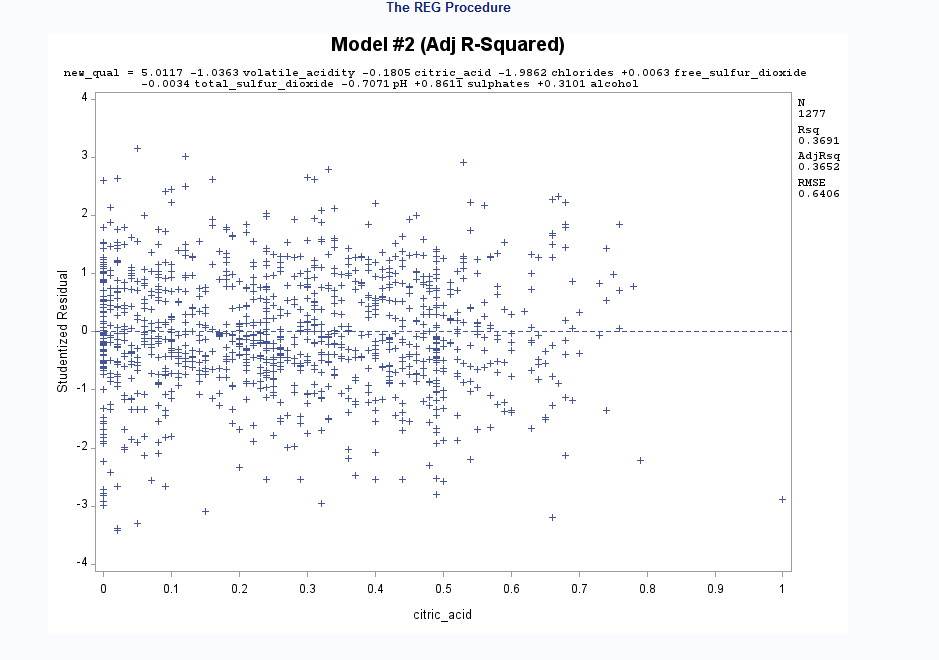
D10



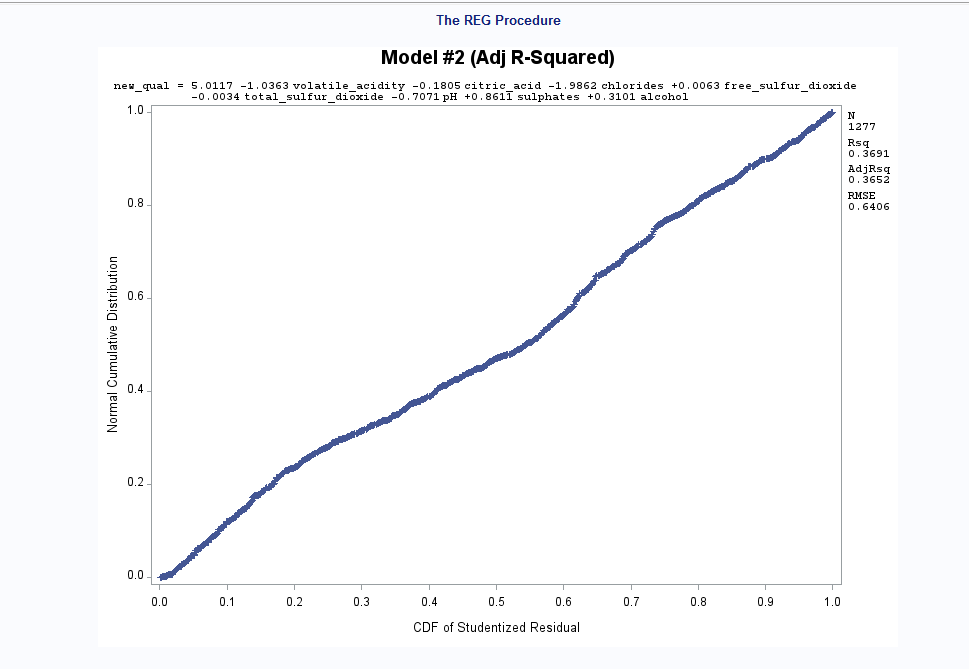
D11



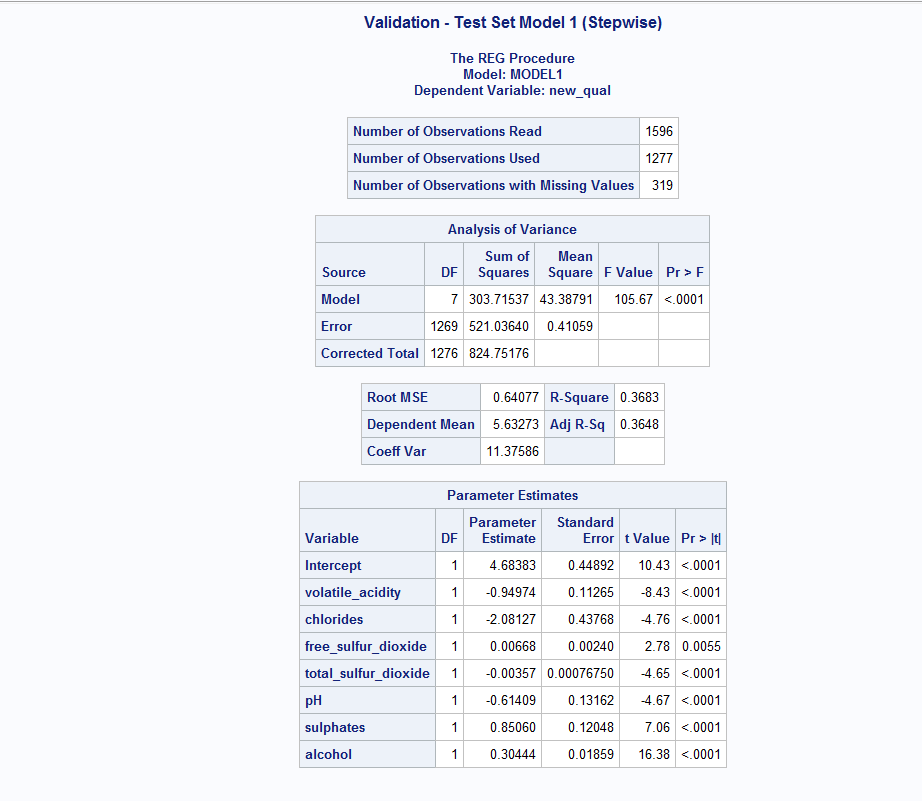
D12



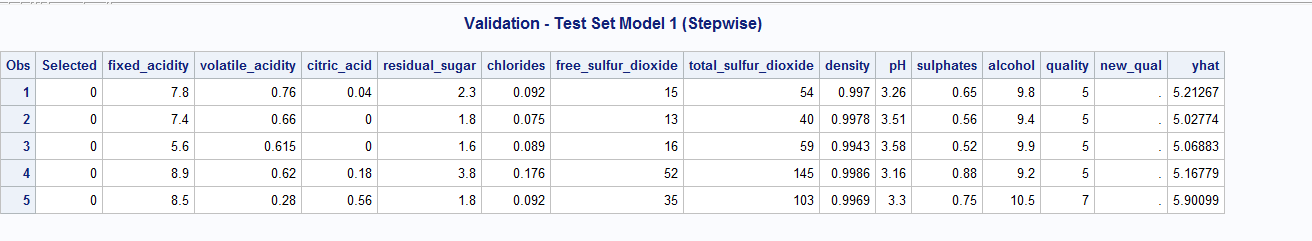
D13



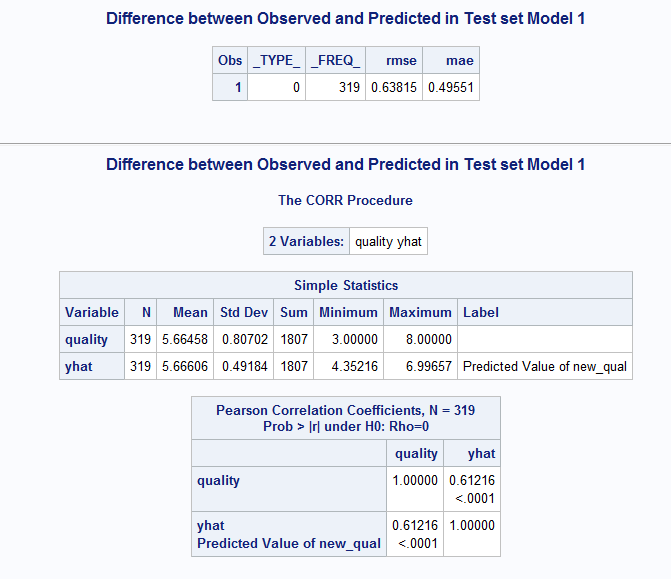
D14



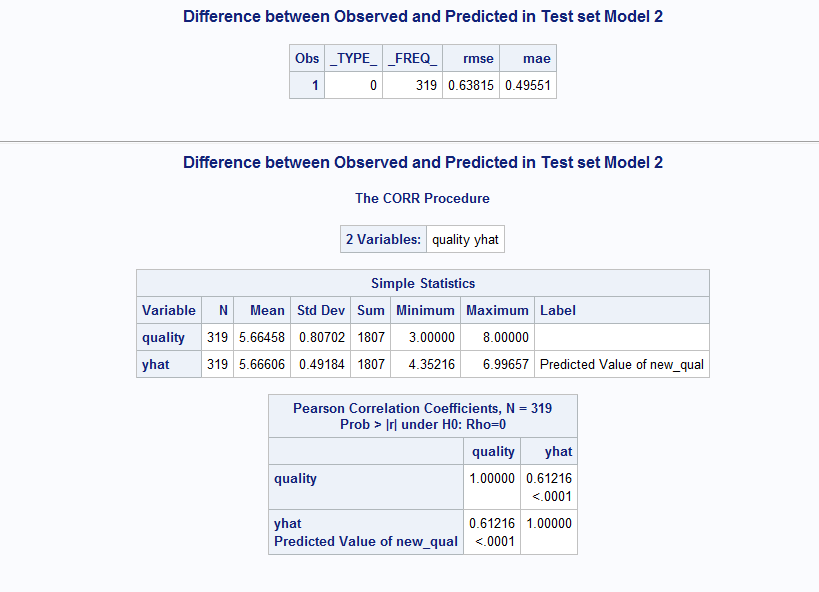
D15



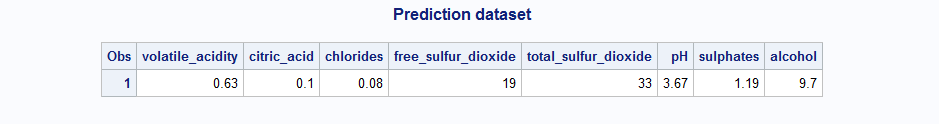
D16



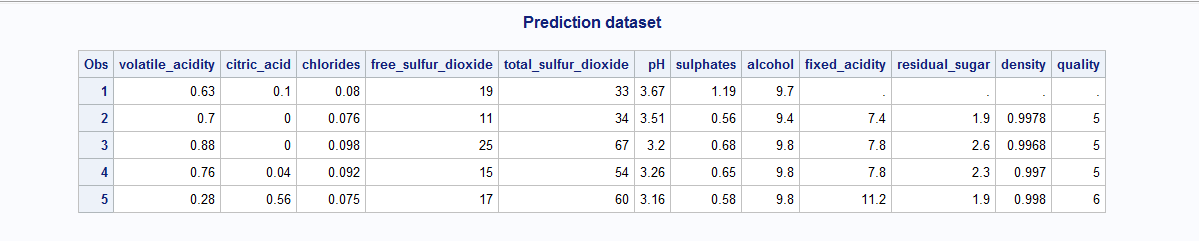
D17



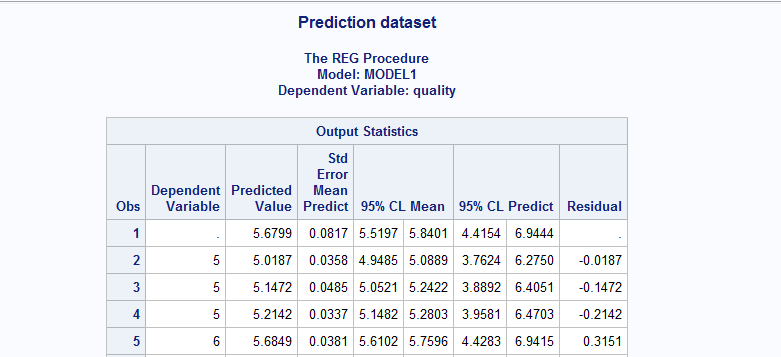
D18



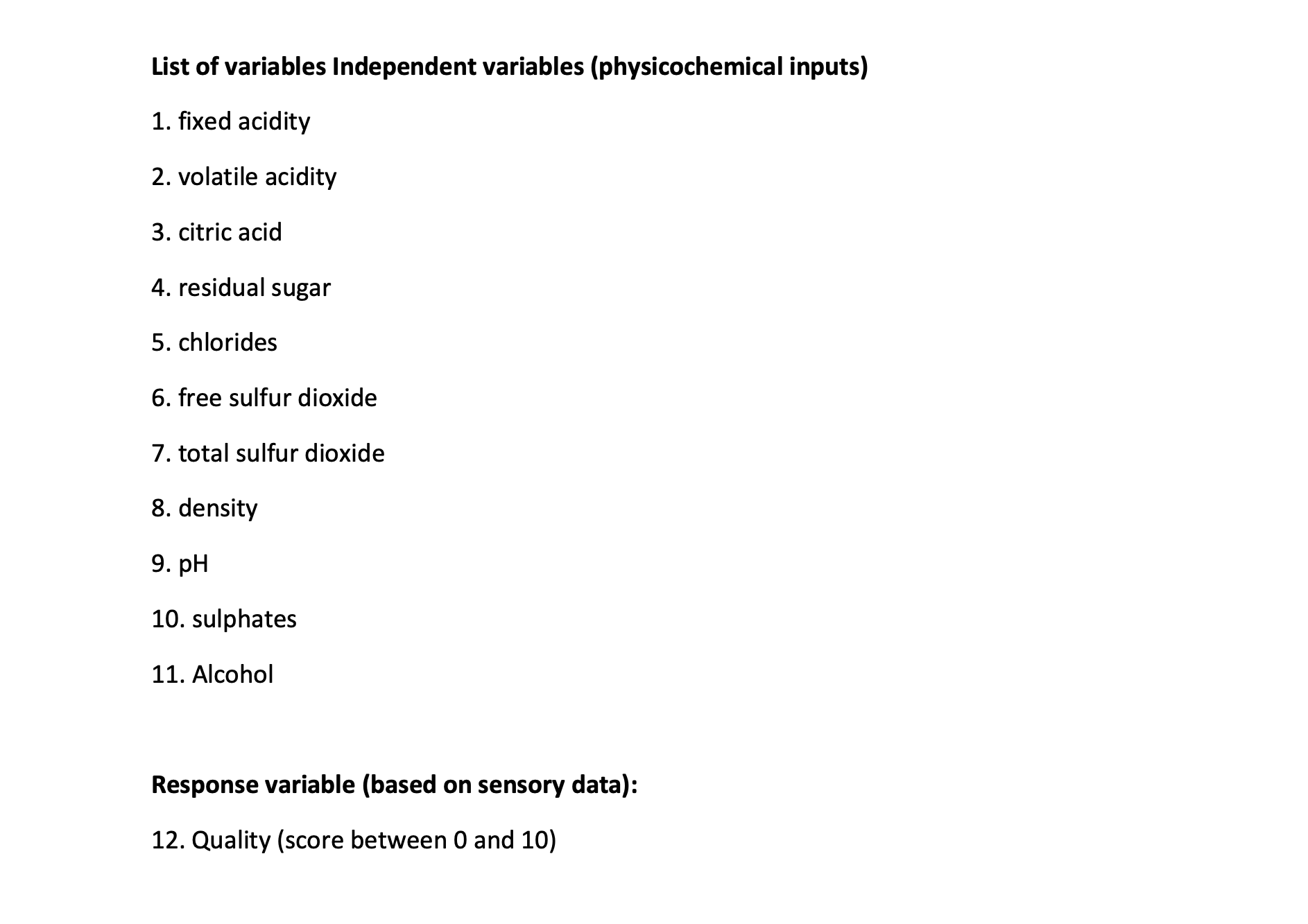
D19



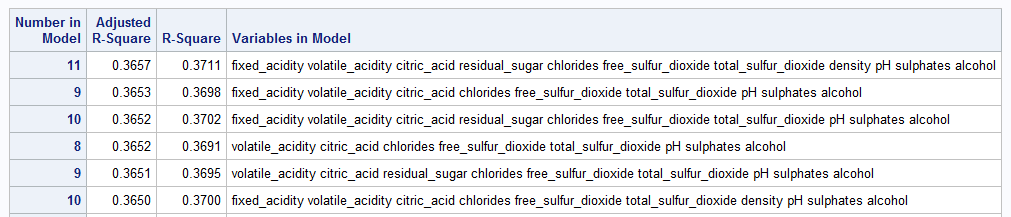
D20



D21



D23



L1

<https://extension.okstate.edu/fact-sheets/understanding-free-sulfur-dioxide-fso2-in-wine.html>

L2

<https://wineserver.ucdavis.edu/industry-info/enology/methods-and-techniques/common-chemical-reagents/citric-acid#:~:text=Citric%20acid%20is%20often%20added,acid%20is%20its%20microbial%20instability>.

L3

<https://wineserver.ucdavis.edu/industry-info/enology/methods-and-techniques/common-chemical-reagents/citric-acid#:~:text=Citric%20acid%20is%20often%20added,acid%20is%20its%20microbial%20instability>.

L4

<https://www.aromadictionary.com/articles/sulfurdioxide_article.html#:~:text=High%20SO2%20can%20also%20render,that%20make%20red%20wine%20red>.

L5

<https://daily.sevenfifty.com/taking-control-of-alcohol-levels-in-wine/#:~:text=Wines%20that%20are%20higher%20in,than%20their%20lower%2Dalcohol%20counterparts>.

<https://www.wineenthusiast.com/basics/drinks-terms-defined/volatile-acidity-wine/>