Build a Wine Quality Multiple Linear Regression Model Using SAS Studio

Didem Bulut Aykurt

MIS500-1 – Foundations of Data Analytics

Colorado State University-Global Campus

Dr. Steve Chung

September 12, 2022

Table of Contents

DATA CLEANING AND MODELLING

- I will extract the data I need from the given source.
- Import the data into SAS.
- Data Preprocessing
- Perform exploratory data analysis.
- Create visualizations to aid exploration.
- Draw my conclusion based on the data.
- Build a regression model to predict the wine quality.
- Selecting the best model based on their respective accuracy.

ATTRIBUTES DESCRIPTION

- Alcohol – The amount of alcohol in wine.
- Volatile acidity – The high levels
- Sulphate - A wine additive that contributes to S02 levels and acts as an antimicrobial and antioxidant.
- Citric acid - It serves as a preservative (small amounts add flavor and freshness to wines)
- Free sulphur dioxide – Is the amount of free + bound forms of SO2
- Density – The more sugar, the more the wine's density.

- Chloride – The amount of salt in the wine.
- Fixed Acidity - The non–volatile acids that do not evaporate readily
- PH - Level of acidity
- Free Sulphur Dioxide – It prevents microbial growth and the oxidation of the wine.
- Residual Sugar - The sugar remains after fermentation stops. Sweetness and sourness must be in perfect balance. (wine >45 grams/liter is sweet).

THE DESIRED OUTCOMES OF THE PROJECT

- Insights of the attributes
- Analysis of the data

WHAT ARE WE TRYING TO ACCOMPLISH?

- What is the number of samples of red wines?
- What is the number of samples of white wine?
- How many columns are there in each dataset?
- Are there any missing values?
- How many duplicate rows are there in the white wine dataset?
- How many duplicate rows in these datasets should be deleted?
- What is the number of unique values of quality in the white wine dataset?
- What is the number of unique values of quality in the red wine dataset?
- Can higher alcohol content wines be rated better?
- Do sweeter wines (with more residual sugar) receive higher ratings?
- Which acidity level is rated the highest on average?

The project goal is to determine the critical attributes that lead to wine quality and to build a model for the prediction of wine quality.

The Wine Quality's Exploratory Data Analysis

I aim to analyze which independent variables affect wine quality with SAS Studio. That way, we know excellent and poor wine products, which helps predict the end of the model. The wine quality dataset; describes concerns about wine quality in the red and white difference between the Portuguese "Vinho Verde" wine. The wine quality data set is available at the UCI Machine Learning repository. Four thousand eight hundred ninety-eight intakes represent aggregated data and 11 attributes(features) with the dependent variable(quality) as follows: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. The variable of quality is the output variable.

Red Wine Dataset's Summary Statistic and Distribution in SAS Studio

+

Figure 3: *The result of* free_sulfur_dioxide's summary statistics in SAS Studio.

15/08/2022, 11:27 Results: Summary Statistics 1

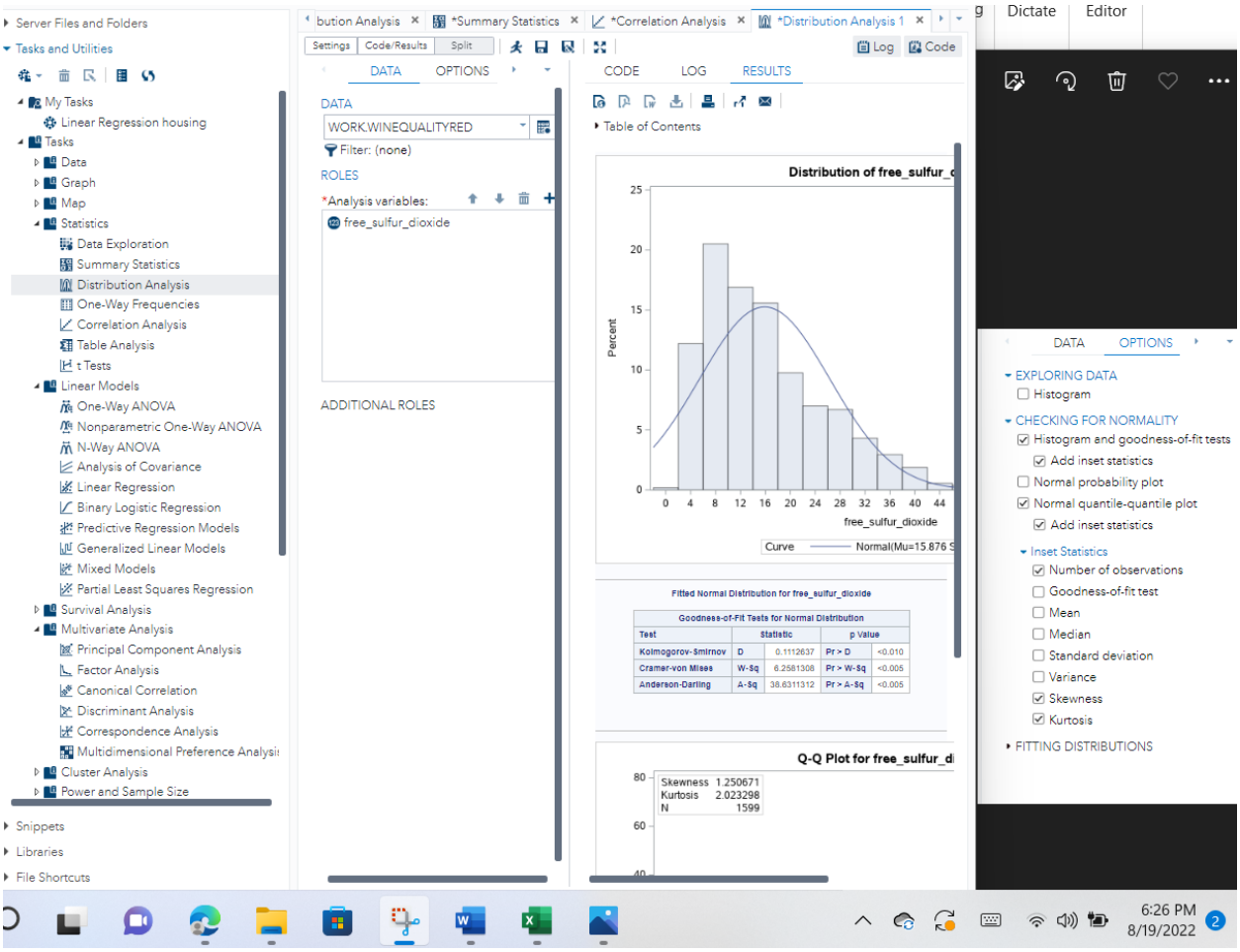
Analysis Variable : free sulfur dioxide free sulfur dioxide				
Mean	Std Dev	Minimum	Maximum	N
15.8749218	10.4601570	1.0000000	72.0000000	1599

Interpretation

The mean and the standard deviation of sulfur dioxide are 15.8749218 and 10.4601570, respectively.

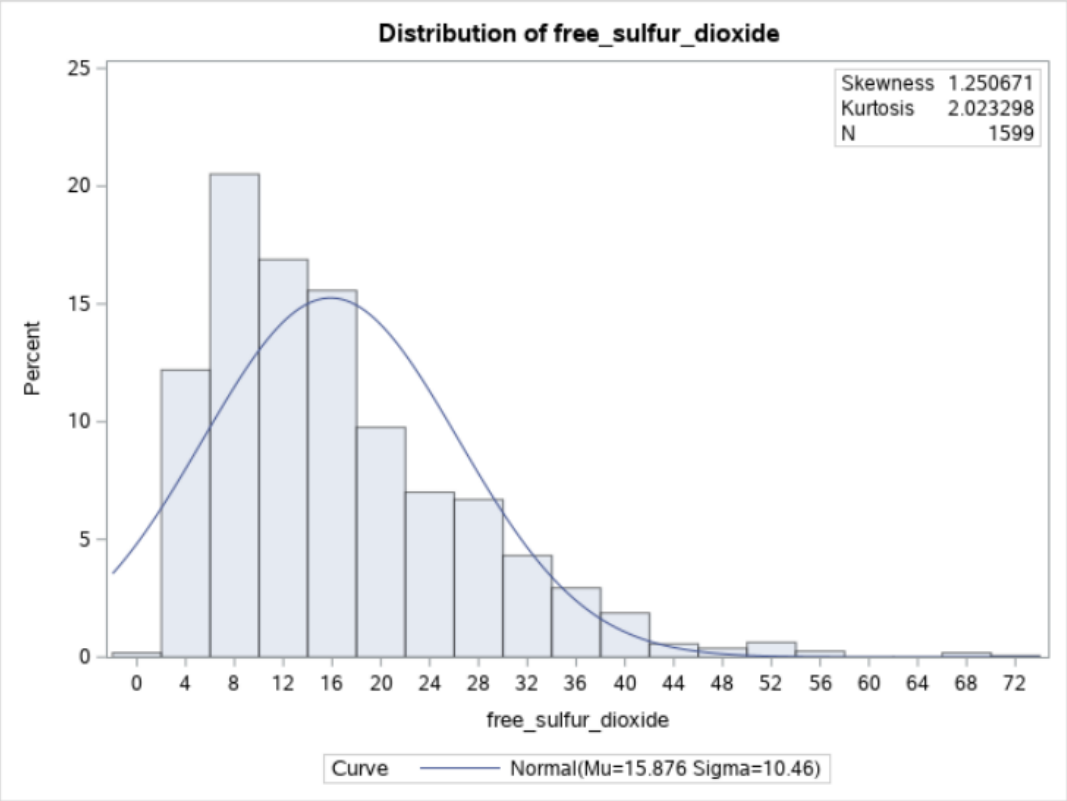
Distribution Analysis for variable residual sugar.

Figure 4: *The step and detail for distribution analyst for "free_sulfur_dioxide" in SAS Studio.*



The desired result is shown below.

Figure 5: *The result of distribution analyst for "free_sulfur_dioxide" in SAS Studio.*



Fitted Normal Distribution for free_sulfur_dioxide

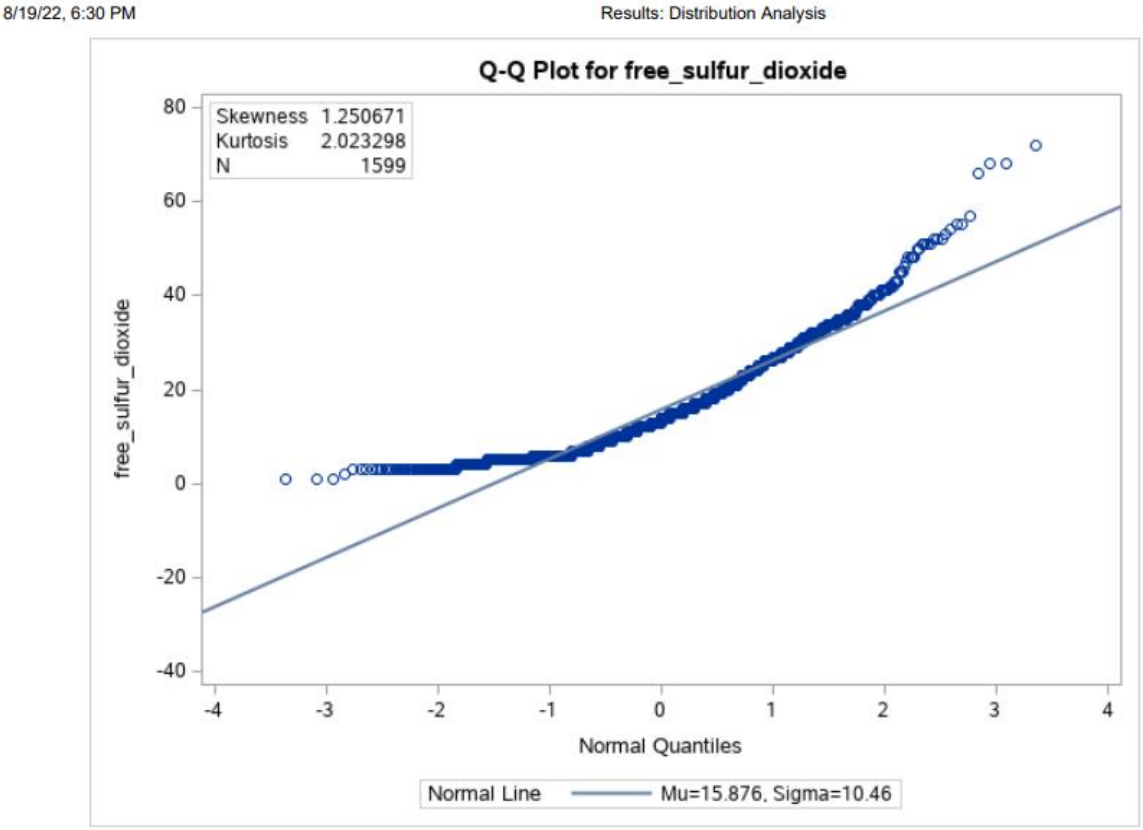
Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.1112637	Pr > D	<0.010
Cramer-von Mises	W-Sq	6.2581308	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	38.6311312	Pr > A-Sq	<0.005

Interpretation

The standard curve shows that the distribution is rightly skewed. A goodness-of-fit test is how well the sample data set fits an entire population with normal distribution and another way to say how target values are related to the independent values in a model. Kolmogorov-Smirnov test applies a large sample of over 2000. The other two tests also have the same reason to use, which helps to

know whether the example of the normal distribution. Our three p-value is lower than 0.05 means the data set is not a normal distribution.

Figure 6: *The result of Q-Q analyst for "free sulfur dioxide" in SAS Studio.*

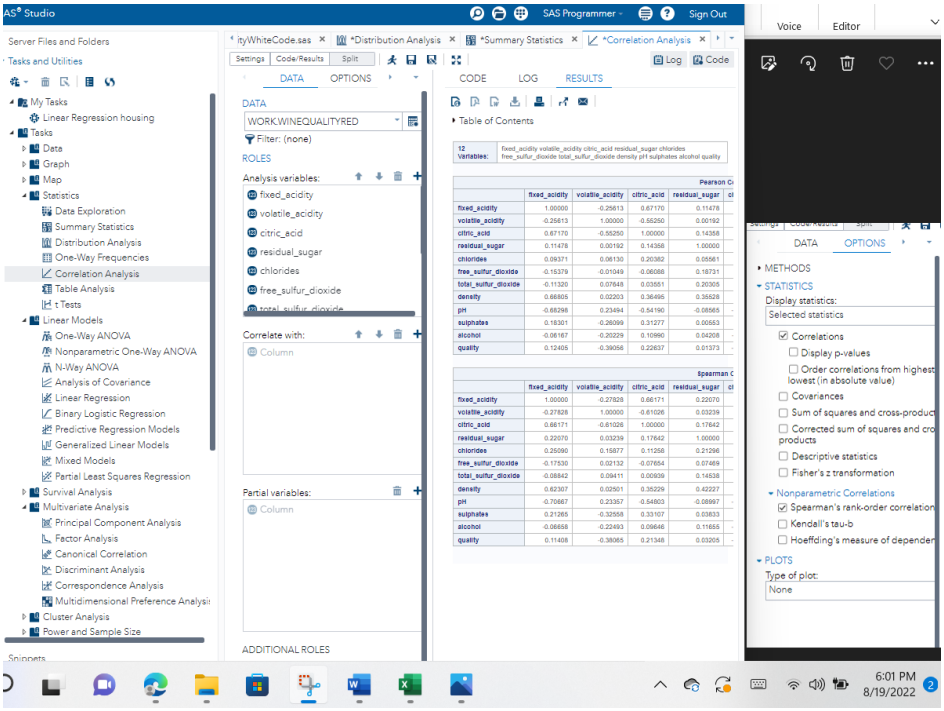


The Q – Q plot above shows that all points do not fall on a straight line. Therefore, the distribution is not normal—also the high end of positive skewness. Kurtosis is more significant than zero means leptokurtic as the dataset has outliers, and this data set outliers affect the linear line.

Correlation Analysis for Red Wine Variables

Each step for correlation analysis in SAS Studio is shown b

Figure 7: *The Correlation Analyst steps in SAS Studio with the wine quality red dataset.*



The required result is given in below.

Figure 8: *The correlation analysis result for all wine quality white variables in SAS Studio.*

8/19/22, 6:05 PM Results: Correlation Analysis

12 Variables:	fixed_acidity volatile_acidity citric_acid residual_sugar chlorides free_sulfur_dioxide total_sulfur_dioxide density pH sulphates alcohol quality

Pearson Correlation Coefficients, N = 1599

	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates	alcohol	quality
fixed_acidity	1.00000	-0.25613	0.67170	0.11478	0.09371	-0.15379	-0.11320	0.66805	-0.68298	0.18301	-0.06167	0.12405
volatile_acidity	-0.25613	1.00000	-0.55250	0.00192	0.06130	-0.01049	0.07648	0.02203	0.23494	-0.26099	-0.20229	-0.39056
citric_acid	0.67170	-0.55250	1.00000	0.14358	0.20382	-0.06088	0.03551	0.36495	-0.54190	0.31277	0.10990	0.22637
residual_sugar	0.11478	0.00192	0.14358	1.00000	0.05561	0.18731	0.20305	0.35528	-0.08565	0.00553	0.04208	0.01373
chlorides	0.09371	0.06130	0.20382	0.05561	1.00000	0.00563	0.04740	0.20063	-0.26503	0.37126	-0.22114	-0.12891
free_sulfur_dioxide	-0.15379	-0.01049	-0.06088	0.18731	0.00563	1.00000	0.66803	-0.02198	0.07029	0.05161	-0.06935	-0.05055
total_sulfur_dioxide	-0.11320	0.07648	0.03551	0.20305	0.04740	0.66803	1.00000	0.07126	-0.06651	0.04292	-0.20567	-0.18511
density	0.66805	0.02203	0.36495	0.35528	0.20063	-0.02198	0.07126	1.00000	-0.34170	0.14851	-0.49618	-0.17492
pH	-0.68298	0.23494	-0.54190	-0.08565	-0.26503	0.07029	-0.06651	-0.34170	1.00000	-0.19665	0.20563	-0.05773
sulphates	0.18301	-0.26099	0.31277	0.00553	0.37126	0.05161	0.04292	0.14851	-0.19665	1.00000	0.09359	0.25140
alcohol	-0.06167	-0.20229	0.10990	0.04208	-0.22114	-0.06935	-0.20567	-0.49618	0.20563	0.09359	1.00000	0.47617
quality	0.12405	-0.39056	0.22637	0.01373	-0.12891	-0.05055	-0.18511	-0.17492	-0.05773	0.25140	0.47617	1.00000

Spearman Correlation Coefficients, N = 1599

	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates	alcohol	quality
fixed_acidity	1.00000	-0.27828	0.66171	0.22070	0.25090	-0.17530	-0.08842	0.62307	-0.70667	0.21265	-0.06658	0.11408
volatile_acidity	-0.27828	1.00000	-0.61026	0.03239	0.15877	0.02132	0.09411	0.02501	0.23357	-0.32558	-0.22493	-0.38065
citric_acid	0.66171	-0.61026	1.00000	0.17642	0.11258	-0.07654	0.00939	0.35229	-0.54803	0.33107	0.09646	0.21348
residual_sugar	0.22070	0.03239	0.17642	1.00000	0.21296	0.07469	0.14538	0.42227	-0.08997	0.03833	0.11655	0.03205
chlorides	0.25090	0.15877	0.11258	0.21296	1.00000	0.00094	0.13004	0.41139	-0.23436	0.02083	-0.28450	-0.18992
free_sulfur_dioxide	-0.17530	0.02132	-0.07654	0.07469	0.00094	1.00000	0.78985	-0.04127	0.11579	0.04581	-0.08137	-0.05686
total_sulfur_dioxide	-0.08842	0.09411	0.00939	0.14538	0.13004	0.78985	1.00000	0.12933	-0.00985	-0.00052	-0.25781	-0.19674
density	0.62307	0.02501	0.35229	0.42227	0.41139	-0.04127	0.12933	1.00000	-0.31206	0.16148	-0.46244	-0.17707
pH	-0.70667	0.23357	-0.54803	-0.08997	-0.23436	0.11579	-0.00985	-0.31206	1.00000	-0.08031	0.17993	-0.04367
sulphates	0.21265	-0.32558	0.33107	0.03833	0.02083	0.04581	-0.00052	0.16148	-0.08031	1.00000	0.20733	0.37706
alcohol	-0.06658	-0.22493	0.09646	0.11655	-0.28450	-0.08137	-0.25781	-0.46244	0.17993	0.20733	1.00000	0.47853
quality	0.11408	-0.38065	0.21348	0.03205	-0.18992	-0.05686	-0.19674	-0.17707	-0.04367	0.37706	0.47853	1.00000

Interpretation

A correlation value higher than 0.5 indicates a strong correlation between the two variables. The correlation value of fewer than 0.5 means there is no strong correlation between the two variables. For instance, the pH and fixed_acidity are strongly negatively correlated with the value of -**0.71,** which will be a 71% negative relation**.**

The total sulfur dioxide and the free sulfur dioxide are correlated with **0.79,** which means a 79% positive correlation. The quality variable doesn't have any strong relationship between all variables for the wine quality red dataset.

White Wine Dataset's Summary Statistic and Distribution in SAS Studio

Select the "total_sulfur_dioxide" variable and obtain the summary statistics. The result is shown in detail below.

Figure 9, *The summary statistic of total_sulfur_dioxide in SAS Studio.*

8/19/22, 6:15 PM Results: Summary Statistics

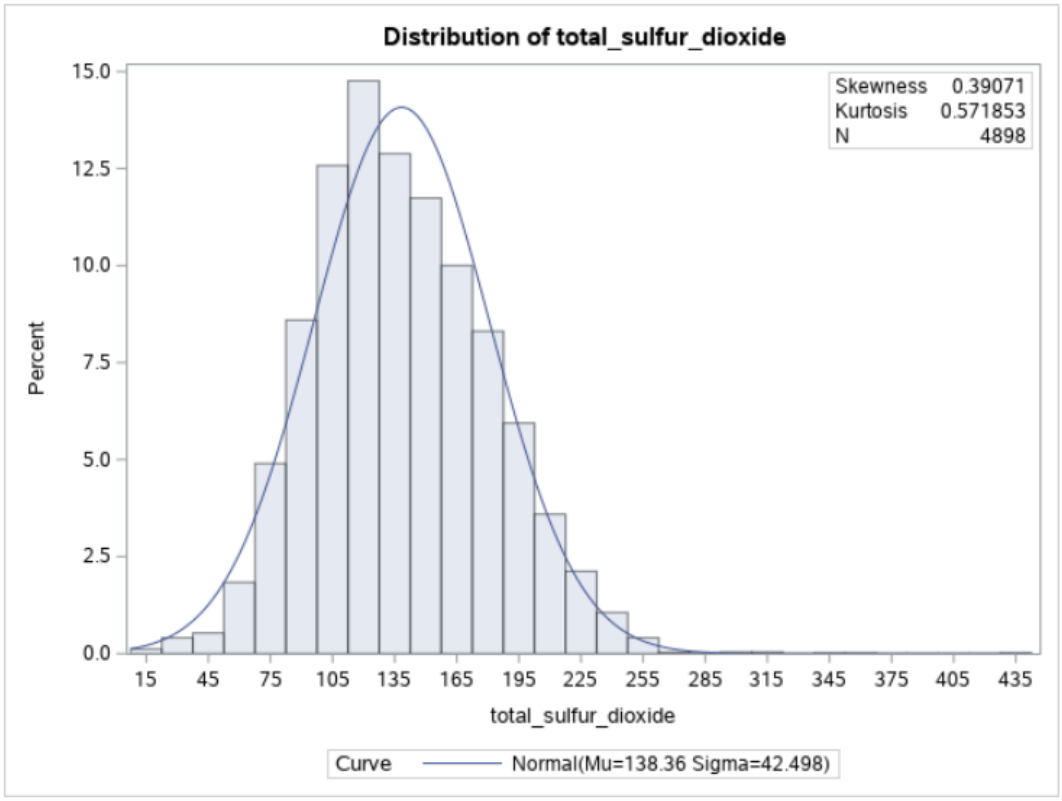
Analysis Variable : total_sulfur_dioxide				
Mean	Std Dev	Minimum	Maximum	N
138.3606574	42.4980646	9.0000000	440.0000000	4898

As a result, shows the range of total_sulfur_dioxide has the highest number. That might predict to affect the quality of the wine. I would like to dive deep into the total_sulfur_dioxide variable.

Distribution Analysis for total sulfur dioxide variable

Analyze the distribution of the "temp" variable in SAS Studio's result given below.

Figure 11, *The resulting distribution analysis is a histogram of the 'RH' variable on SAS studio on SAS studio.*

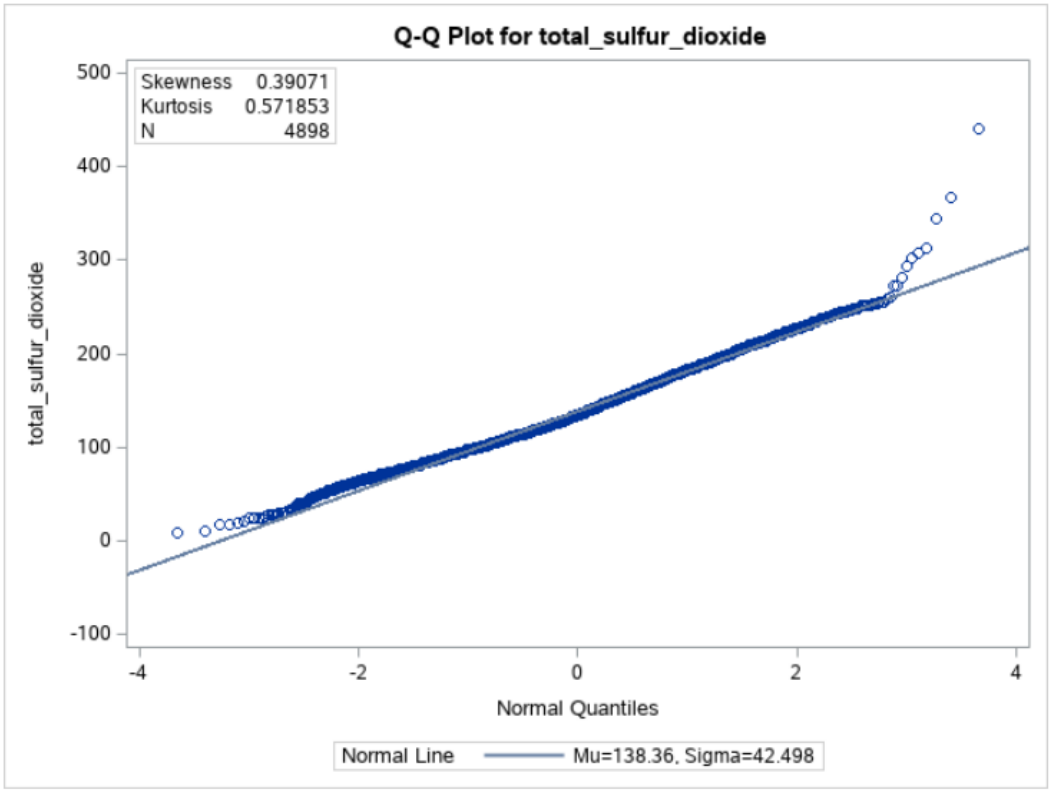


Fitted Normal Distribution for total_sulfur_dioxide

Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.0446500	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	2.1314344	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	11.9389793	**Pr > A-Sq**	<0.005

The above data fit the standard curve. Thus, the distribution looks to follow a normal distribution. But skewness has a positive number of 0.39, which shows positive skewness. Kurtosis>0 means high kurtosis is leptokurtic. The dataset would have outliers. The goodness-of-fit test's all p-value results lower than 0.05 means the dataset is not a normal distribution.

Figure 12, *The resulting distribution analysis is the Norma quantile-quantile plot of the 'temp' variable on SAS studio.*



The above data points fall on a straight line; thus, the Q – Q plot fits the standard curve. We discussed that the previous kurtosis cart shows the dataset might have outliers, then The Q-Q chart proves it. There are a few outliers as they aren't affected by line data.

Correlation Analysis for White Wine Variables

The desired result of correlation analysis in SAS Studio with the white wine dataset is shown below.

Figure 14, *The correlation analysis result for all wine quality white variables in SAS Studio.*

8/19/22, 3:23 PM Results: Correlation Analysis

12 Variables:	fixed_acidity volatile_acidity citric_acid residual_sugar chlorides free_sulfur_dioxide total_sulfur_dioxide density pH sulphates alcohol quality

Pearson Correlation Coefficients, N = 4898

	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates	alcohol	quality
fixed_acidity	1.00000	-0.02270	0.28918	0.08902	0.02309	-0.04940	0.09107	0.26533	-0.42586	-0.01714	-0.12088	-0.11366
volatile_acidity	-0.02270	1.00000	-0.14947	0.06429	0.07051	-0.09701	0.08926	0.02711	-0.03192	-0.03573	0.06772	-0.19472
citric_acid	0.28918	-0.14947	1.00000	0.09421	0.11436	0.09408	0.12113	0.14950	-0.16375	0.06233	-0.07573	-0.00921
residual_sugar	0.08902	0.06429	0.09421	1.00000	0.08868	0.29910	0.40144	0.83897	-0.19413	-0.02666	-0.45063	-0.09758
chlorides	0.02309	0.07051	0.11436	0.08868	1.00000	0.10139	0.19891	0.25721	-0.09044	0.01676	-0.36019	-0.20993
free_sulfur_dioxide	-0.04940	-0.09701	0.09408	0.29910	0.10139	1.00000	0.61550	0.29421	-0.00062	0.05922	-0.25010	0.00816
total_sulfur_dioxide	0.09107	0.08926	0.12113	0.40144	0.19891	0.61550	1.00000	0.52988	0.00232	0.13456	-0.44889	-0.17474
density	0.26533	0.02711	0.14950	0.83897	0.25721	0.29421	0.52988	1.00000	-0.09359	0.07449	-0.78014	-0.30712
pH	-0.42586	-0.03192	-0.16375	-0.19413	-0.09044	-0.00062	0.00232	-0.09359	1.00000	0.15595	0.12143	0.09943
sulphates	-0.01714	-0.03573	0.06233	-0.02666	0.01676	0.05922	0.13456	0.07449	0.15595	1.00000	-0.01743	0.05368
alcohol	-0.12088	0.06772	-0.07573	-0.45063	-0.36019	-0.25010	-0.44889	-0.78014	0.12143	-0.01743	1.00000	0.43557
quality	-0.11366	-0.19472	-0.00921	-0.09758	-0.20993	0.00816	-0.17474	-0.30712	0.09943	0.05368	0.43557	1.00000

Spearman Correlation Coefficients, N = 4898

	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates	alcohol	quality
fixed_acidity	1.00000	-0.04287	0.29788	0.10672	0.09469	-0.02454	0.11265	0.27003	-0.41834	-0.01324	-0.10683	-0.08449
volatile_acidity	-0.04287	1.00000	-0.15041	0.10863	-0.00493	-0.08121	0.11761	0.01012	-0.04520	-0.01690	0.03397	-0.19656
citric_acid	0.29788	-0.15041	1.00000	0.02462	0.03266	0.08831	0.09322	0.09143	-0.14619	0.07977	-0.02917	0.01833
residual_sugar	0.10672	0.10863	0.02462	1.00000	0.22784	0.34611	0.43125	0.78036	-0.18003	-0.00384	-0.44526	-0.08207
chlorides	0.09469	-0.00493	0.03266	0.22784	1.00000	0.16705	0.37524	0.50830	-0.05401	0.09393	-0.57081	-0.31449
free_sulfur_dioxide	-0.02454	-0.08121	0.08831	0.34611	0.16705	1.00000	0.61862	0.32782	-0.00627	0.05225	-0.27257	0.02371
total_sulfur_dioxide	0.11265	0.11761	0.09322	0.43125	0.37524	0.61862	1.00000	0.56382	-0.01183	0.15782	-0.47662	-0.19668
density	0.27003	0.01012	0.09143	0.78036	0.50830	0.32782	0.56382	1.00000	-0.11006	0.09508	-0.82186	-0.34835
pH	-0.41834	-0.04520	-0.14619	-0.18003	-0.05401	-0.00627	-0.01183	-0.11006	1.00000	0.14024	0.14886	0.10936
sulphates	-0.01324	-0.01690	0.07977	-0.00384	0.09393	0.05225	0.15782	0.09508	0.14024	1.00000	-0.04487	0.03332
alcohol	-0.10683	0.03397	-0.02917	-0.44526	-0.57081	-0.27257	-0.47662	-0.82186	0.14886	-0.04487	1.00000	0.44037
quality	-0.08449	-0.19656	0.01833	-0.08207	-0.31449	0.02371	-0.19668	-0.34835	0.10936	0.03332	0.44037	1.00000

The correlation coefficient between the variable total sulfur dioxide and free sulfur dioxide is highly correlated at 91%. The variable density and the residual sugar are highly correlated at 83%.

MULTILINEAR REGRESSION MODEL

The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. These datasets can be viewed as a regression task. A linear regression model could help detect excellent or poor wines predictor. The project is focused on observing the wine quality rates. The quality of the wine is a dependent variable, and the independent variables are fixed_acidity, volatile_acidity, citric_acid, residual_sugar, chlorides, free_sulfur_dioxide, total_sulfur_dioxide,

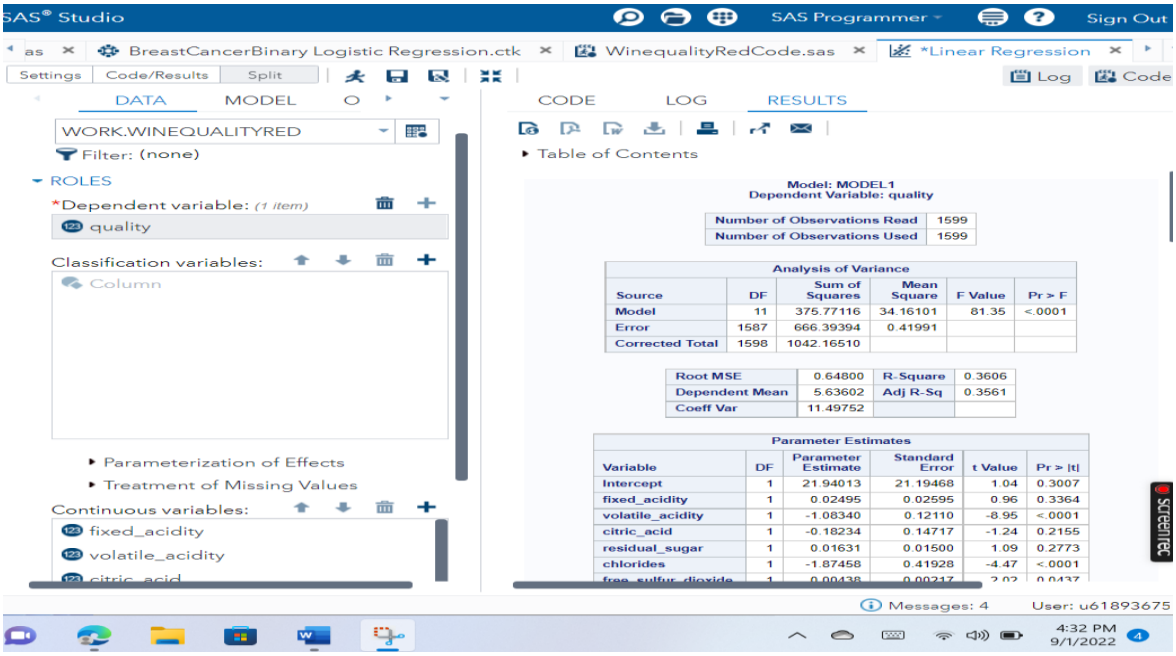
density, pH, sulphates, and alcohol. I will use two linear regression and two linear models with the Forward Selection technique to analyze the red and white datasets.

Multilinear Regression Model to the "winequalityRed" dataset in SAS Studio.

Linear regression was used to identify the possible predictors of the dependent variable quality out of the following dependent variables. We will show relationships between dependent and continuous variables at each step. If the model fits this dataset, we can find predicted variables.

I created a linear regression model and applied all independent variables in the "winequalityRed" dataset. The first step is to click "Tasks and Utilities" on the main menu and then double click "Linear Regression" under the "Linear Model" in SAS Studio. It will pop up a new Linear regression page. The work panel has three options. One is the "DATA" organizer to choose data "WORK.WINEQUALITYRED" next to the dependent variable "quality" and the eleven continuous variables. The next second is the "Model," which selects all constant variables. The final part is "Option" to create diagnostic plots and residuals for each variable.

Figure 1
The multiLinear regression model's step of winequalityRed dataset in SAS Studio.



The result for the multilinear regression model of winequalityRed data set in SAS Studio.

The first result table in Figure 2 has statistical detail; the observation read and used data number is 1599, which means no missing value. The ANOVA table has great points like R-square, P-value, F-value, Sum of squares, and RMSE.

Look at each result to see how our data fit a linear and normal distribution. **SS**'s best result is zero, and the case result of 375. A higher number to tell the winequalityRed dataset fits the linear regression model. The one-way ANOVA measures the F value, how a group of variables is jointly significant, and is used to decide to support or reject the null hypothesis. The result is the F value of 81.35, which is not a small enough number to say all variables are statistically significant. P-value has a lower number of 0.001, meaning there is much variability in my target. R-square tells us how strong the relation between two variables value is from 0 to 100; the highest number means a great fitness model. The result of 0.36 is 36%, pretty low to apply all variables. Root mean squared error (RMSE) helps us see how wrong the model's predictions are compared to actual

observed values. So, a high RMSE is 'bad,' and a low RMSE is 'good.' The result's RMSE is 0.64, which is standard and means the predictors fit the model.

The following table is Parameter Estimates, with results of t-value and p-value for all independent variables. P-value tells us the model is a perfectly supported dataset, so six variables in my target result in a p-value lower than 0.05. However, if the p-value is higher than 0.05, that variable would not include in the model. A t-value of zero best results mean sample results equal the null hypothesis.

Thus, our model should include volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, pH, sulphates, and alcohol.

Figure 2

The result of the ANOVA table of winequalityRed in SAS Studio.

9/1/22, 4:20 PM Results: Linear Regression

Model: MODEL1
Dependent Variable: quality

Number of Observations Read	1599
Number of Observations Used	1599

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	375.77116	34.16101	81.35	<.0001
Error	1587	666.39394	0.41991		
Corrected Total	1598	1042.16510			

Root MSE	0.64800	R-Square	0.3606
Dependent Mean	5.63602	Adj R-Sq	0.3561
Coeff Var	11.49752		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	21.94013	21.19468	1.04	0.3007
fixed_acidity	1	0.02495	0.02595	0.96	0.3364
volatile_acidity	1	-1.08340	0.12110	-8.95	<.0001
citric_acid	1	-0.18234	0.14717	-1.24	0.2155
residual_sugar	1	0.01631	0.01500	1.09	0.2773
chlorides	1	-1.87458	0.41928	-4.47	<.0001
free_sulfur_dioxide	1	0.00438	0.00217	2.02	0.0437
total_sulfur_dioxide	1	-0.00327	0.00072903	-4.49	<.0001
density	1	-17.85480	21.63322	-0.83	0.4093
pH	1	-0.41398	0.19160	-2.16	0.0309
sulphates	1	0.91629	0.11434	8.01	<.0001
alcohol	1	0.27620	0.02648	10.43	<.0001

The following graph, Observed by Predicted quality in Figure 3, tells the dependent variable and independent variable don't have a linear connection, and the data points do not lie close to the fit residual=0 line, which means this model is not an excellent fit dataset.

Follow plot chart is the Fit Diagnostics for quality has many graphs on it. Let's look at each of the detail; the residual of the distribution plots (1,2,3) have a heavy-tailed residual, which means the linear regression model is not a good choice because all the variables are not distributed homogeneously around the residual line or are not randomly scattered against the predicted value as the residual should equal zero but not in this case. That means our model's not picking up all the signals of the dataset. The residual-fit spread plot helps to compare the spread of the fit and the spread of the residual. Our chart shows that the right property (residual) is taller than the left plot(fit-mean), as we can say the model can't explain all the variations. The residual chart has skew distribution if the proper distribution (residual plot) is skewed. The quantile plot of residuals shows that the residual and quantile relationship is close to saying linear, as you can see, has outliers. The cook's distance help to find influential outliers in predictor variables which means knowing which points negatively affect our model; that point should be more than 1.0. The Cook's D plot has a few moments but no more than one Cook's D result, which means not worrying about outliers or negative affect.

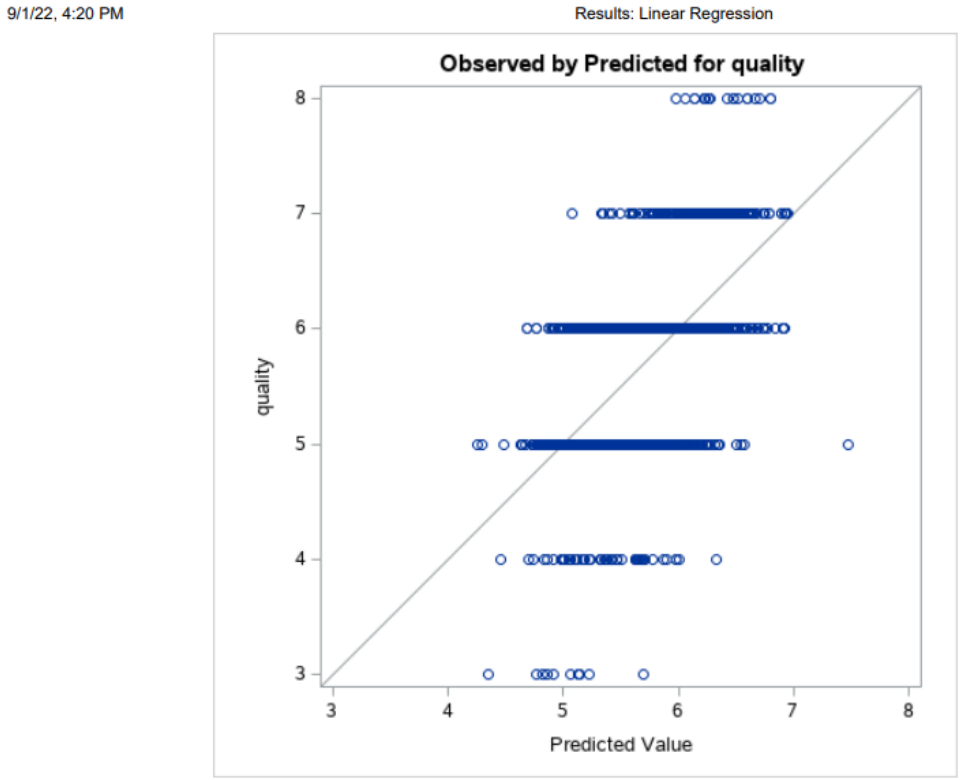
In the final graph, the Residual by Regressors plots for each independent variable, All the variable chart has noisy data. It is hard to say something for regression of all variables; then, I used LOESS Smooth (added to complex code plots= residuals(smooth) that help to filter the noise and see all potential regression signals. All the variables' plots aren't smoothly distributed vertically, as they have outliers, except pH seems to be a quadratic relationship. We need more stories; it has many outliers, around 3.2. That point has a peak at approximately a point of zero.

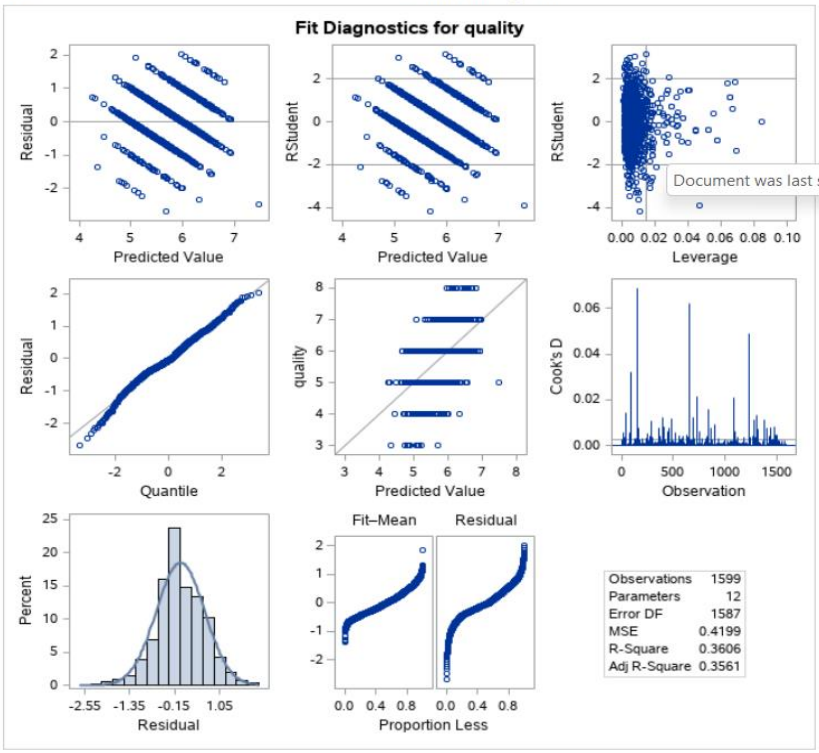
Thus, all the results do not support a linear regression model that is dependent and independent and should be linear, independency, normal distribution, and equal error. That is why the

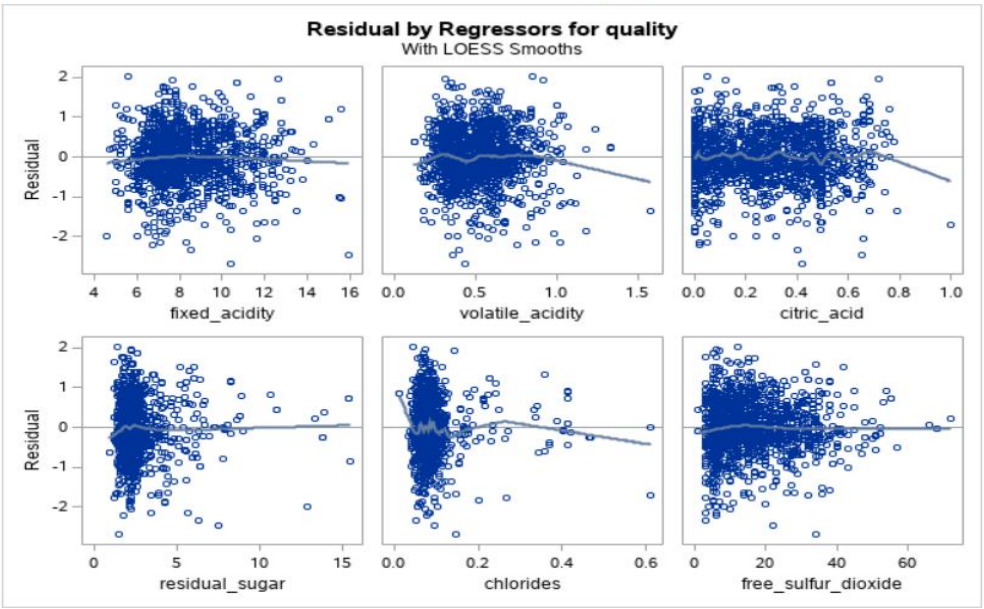
multilinear model does not support the winequalityRed dataset and might add more variables to help improve model results or apply a nonlinear model.

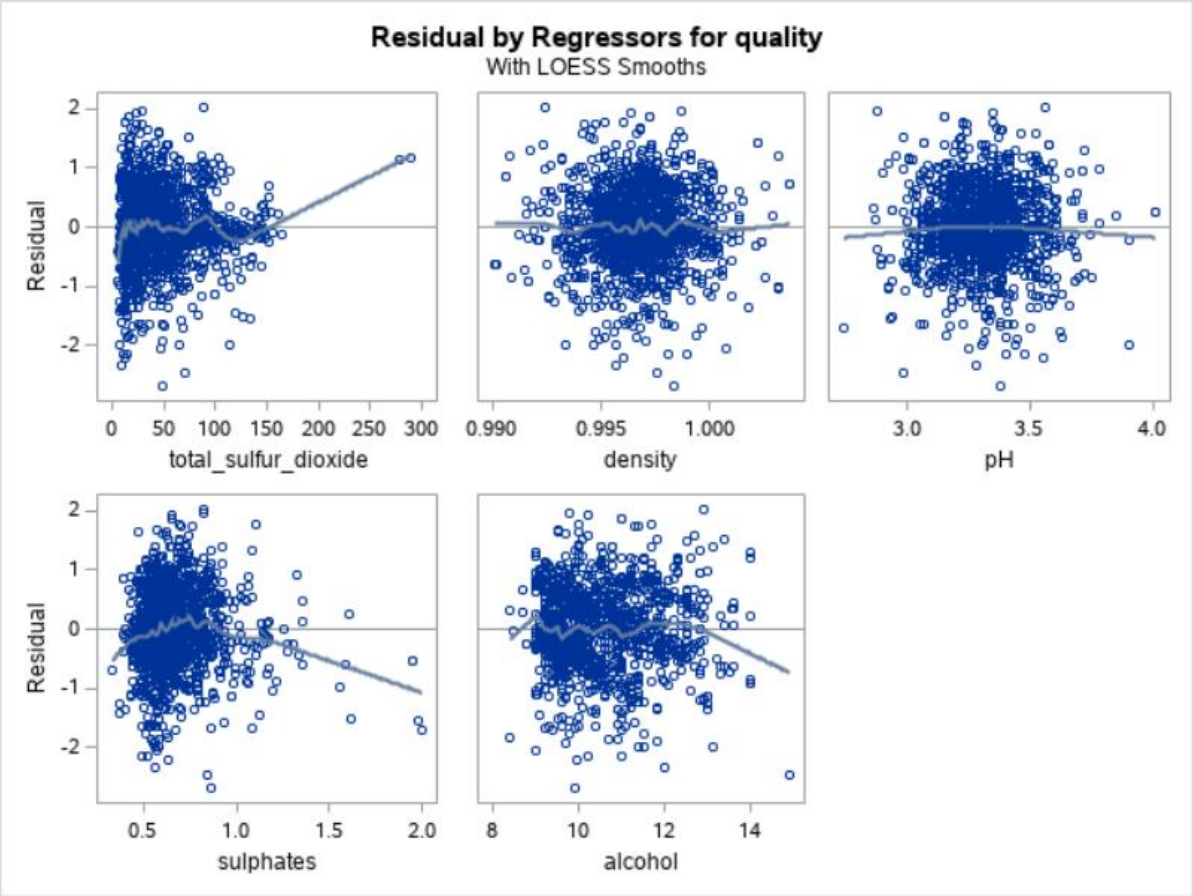
Figure 3

The multilinear regression model for the winequalityRed dataset in SAS Studio results.




Fit Diagnostics for quality
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Document was last saved: 58
Residual
Quantile
quality
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 1599
Parameters 12
Error DF 1587
MSE 0.4199
R-Square 0.3606
Adj R-Square 0.3561


Residual by Regressors for quality
With LOESS Smooths
Residual
fixed_acidity
volatile_acidity
citric_acid
Residual
residual_sugar
chlorides
free_sulfur_dioxide



Forward Selection Technique by "winequalityRed" into SAS Studio

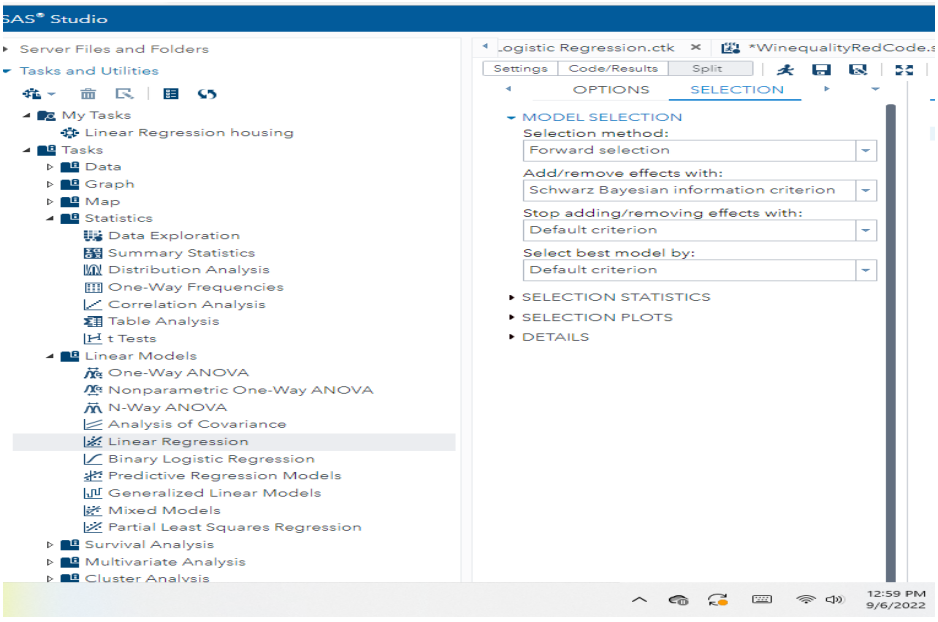
In forward selection, feature variables are examined one at a time, so we can look at all variables to find the best one to see if it can go into the model and be put in. Then look at the remaining ones, find the best one, and put it in possibly. The analyst sets a stopping rule (several such as p-value) that acts as a hurdle the variable must overcome to be allowed into the model. The added variable must reduce error significantly. The feature variable that most reduces model error (SSE) is chosen so long as it passes the stopping rule. Once a variable is in the model, it stays. It is never removed. Then the repeats until all variables are in the model or no variable passes the stopping rule. Suppressor variables are strange. They correlate with parts of the variable but not the target. They eat up some of the non-relevant errors in those variables, making the model more robust, and

forward selection can miss those. There are also complimentary variables that are negatively correlated with each other but together explain the target variable very well, and forward selection the way it is constructed can also miss those relationships.

Figure 4 has details; first, click “Task and Utilities” in the main list, then select “Linear Regression” under “Linear Model,” then select “WORK.WINEQUALITYRED” next to the dependent variable “quality” and the eleven continuous variables. The next second one is the “Model,” which sets all constant variables. The final part is the “Selection” tab to tick Forward selection. Add hard code LOESS (plots= residuals(smooth)) to eliminate data noise to see quick distribution.

Figure 4

The Forward Section technique in multiLinear regression model’s step of winequalityRed dataset in SAS Studio.



The result for the multilinear regression model with Forward Selection technique for winequalityRed data set in SAS Studio.

The result obtained is shown below in figure 5; the first two tables have data detail in the model is total variable is 12 and observations are 1599; the next one is Forward selection Summary, as we found the previous model in seven variables have a result of lower than 0.05 p-value and the forward selection technique selected, filtered and added model those six variables are alcohol, volatile acidity, sulfates, total_sulfur_dioxode, chlorides, and pH as a more statistically significant. The forward selection technique sets a stopping rule, so it didn't add free sulfur dioxide to the model. The Stop Detail table shows SBC has a high negative number increase and that negative probability correlation, which is why the stopping rule didn't add the model.

The following graph, Fit Criteria for Quality, results from AIC (Akaike Information Criterion), AICC, Adjusted R-squared, and SBC for six independent variables that help to compare two model results to choose which model is the best for the dataset. The following table stops the rule. The R-square is low at 35% common percentage relation between dependent and continuous variables, and SS is high than zero, so the dataset does not perfectly fit the model. The RMSE is the low point of 0.6 which shows how good the model's predictions are, but that is not enough to tell the model fits the dataset. The t-value's zero is the best result to tell mean sample results equal to the null hypothesis that also describes the normal distribution. Additionally, the development of the t-value helps to know how the sample test aimed the hypothesis testing to give the population result safe. If the t-value is zero, we can reject the null hypothesis as in our model; all variable results do not equal zero. Thus, we can't deny the null hypothesis.

Predicted observes the fallow plot. As we can see, any data points are not close to the diagonal line, so the model does not expect the winequalityRed dataset. Let's look at the other plots, which can help us validate our assumptions, the Fit Diagnostics for quality graph has a multi-chart. One

of the Residual shows predicted values on the x-axis and residuals on the y-axis, and each point's distance from the line (zero) tells how wrong the prediction for the model is. Thus, the first two chard's data points do not homogamically stay around the line at zero. As a result, all predictors are not suitable for this model. The Quantile chart shows the most point on the line, but we can say the linear and the bottom and top have outliers far from the line. There are outliers in the dataset. As we can see, Cook's D has a few high peaks. One is the stay 0.1. That point could affect the model. **T**he residual-fit spread plot helps to compare the spread of the fit and the spread of the residual. Our chart shows that the right property (residual) is taller than the left plot(fit-mean), as we can say the model can't explain all the variations. The residual chart isn't standard if the proper distribution (residual plot) is skewed. The Residual by Regressors for quality with LOESS Smooth's has all been predicted on the x-axis and residual y-axis. The model has an all predicter by the line so bad that means the model doesn't accurately represent the relationship between predict and quality of wine variable.

Thus, each chart, graph, and table tell that the winequalityRed dataset doesn't fit the multilinear regression model. Removing outliers, adding more variables, or using a nonlinear model can give the dataset a better result. The linear model has to have four rules that need to meet to perform a linear regression. Independency as random scatter our data point not support that assumption, normal distribution, and errors terms are the distance, or the value between the predicted value that comes from our line and actual value and equal error variance that means don't have any patterns as our model result shows many patterns on the Residual by regressors for quality plots. The result indicates that winequalityRed is not fit the linear regression model.

Figure 5

The result of the Forward Section technique for the "winequaliteRed" dataset in SAS Studio.

Data Set	WORK.WINEQUALITYRED
Dependent Variable	quality
Selection Method	Forward
Select Criterion	SBC
Stop Criterion	SBC
Effect Hierarchy Enforced	None

Number of Observations Read	1599
Number of Observations Used	1599

Dimensions	
Number of Effects	12
Number of Parameters	12

Forward Selection Summary

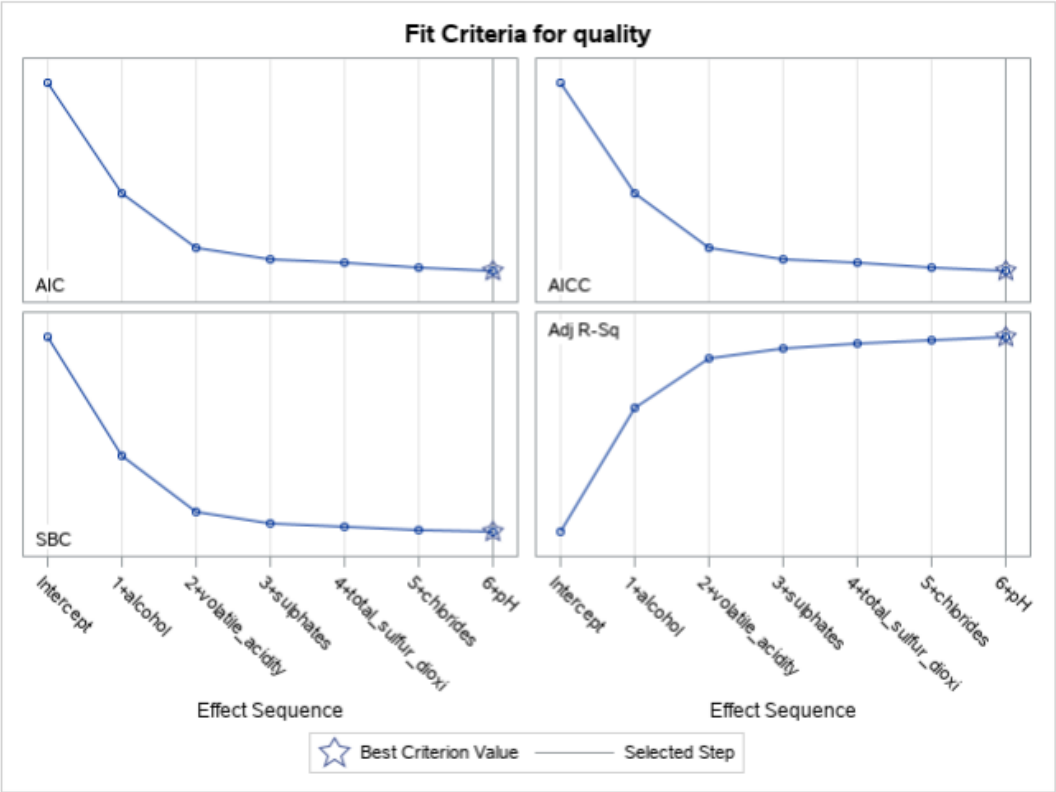
Step	Effect Entered	Number Effects In	SBC
0	Intercept	1	-677.1197
1	alcohol	2	-1080.8977
2	volatile_acidity	3	-1272.0065
3	sulphates	4	-1309.4886
4	total_sulfur_dioxide	5	-1321.2107
5	chlorides	6	-1332.7359
6	pH	7	-1339.4245*

* Optimal Value of Criterion

Selection stopped at a local minimum of the SBC criterion.

Stop Details

Candidate For	Effect	Candidate SBC		Compare SBC
Entry	free_sulfur_dioxide	-1337.8191	>	-1339.4245



Selected Model

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	372.23410	62.03902	147.43	<.0001
Error	1592	669.93100	0.42081		
Corrected Total	1598	1042.16510			

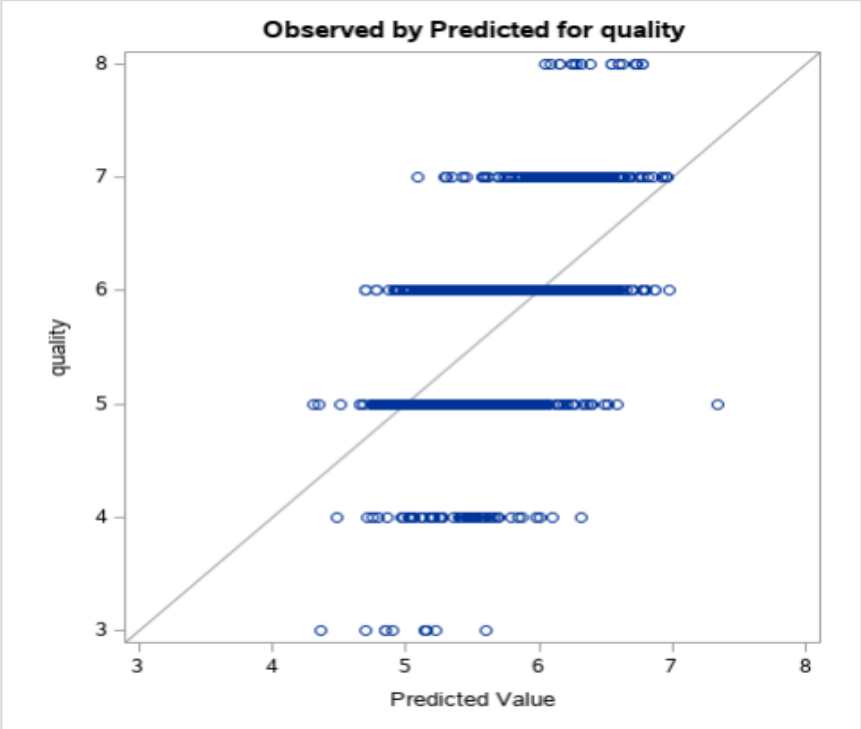
Root MSE	0.64870
Dependent Mean	5.63602
R-Square	0.3572
Adj R-Sq	0.3548
AIC	223.93558
AICC	224.02615
SBC	-1339.42448

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|

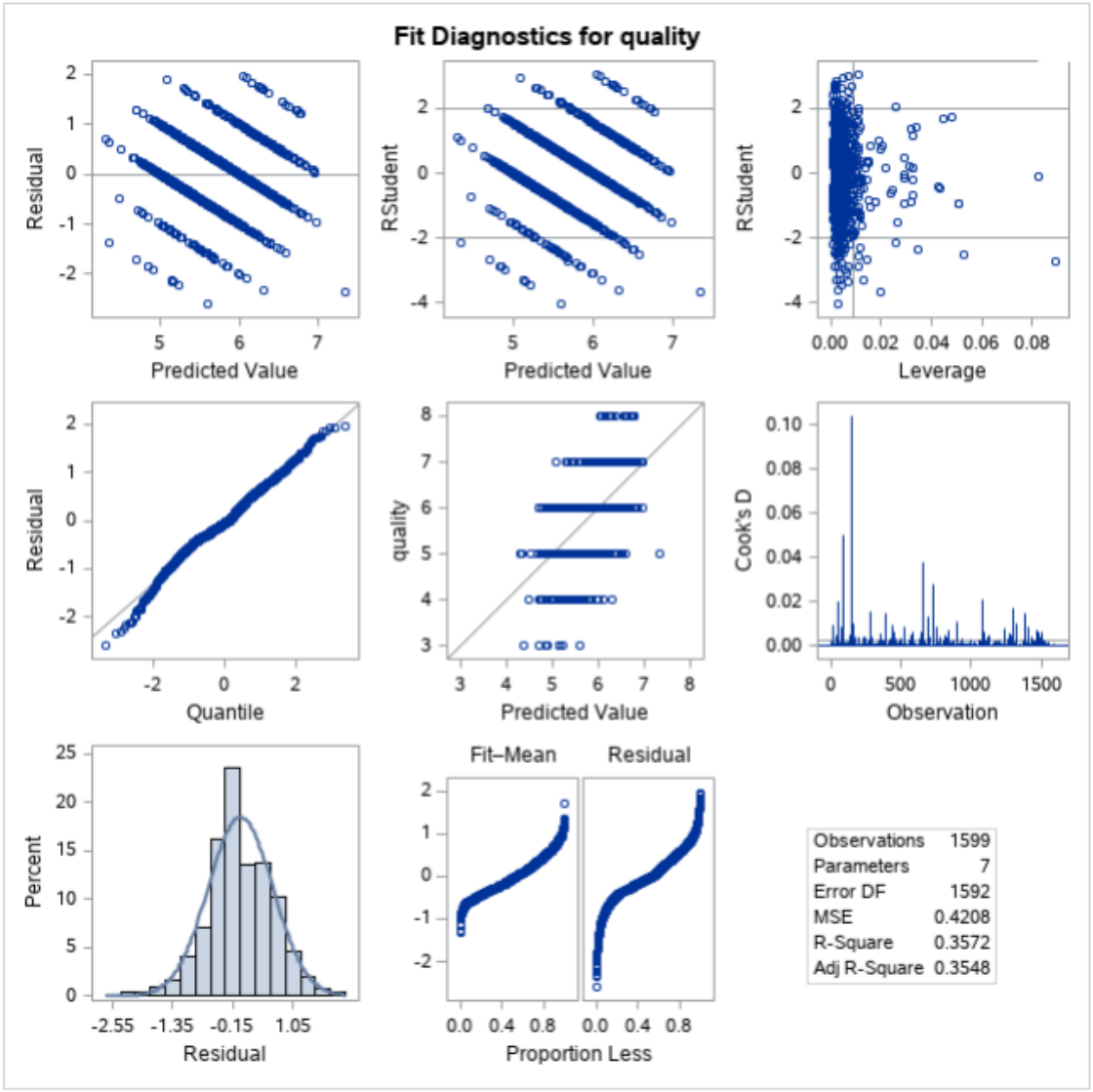
Parameter Estimates

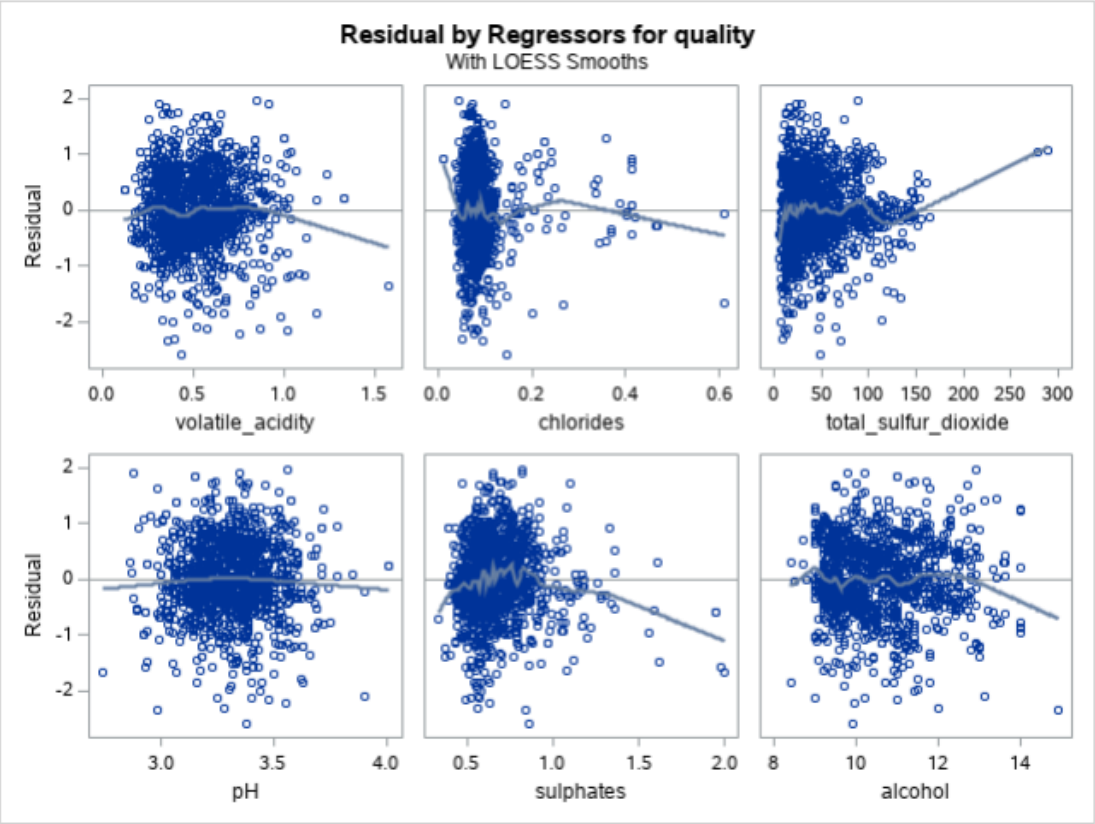
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.295772	0.399562	10.75	<.0001
volatile_acidity	1	-1.038192	0.100427	-10.34	<.0001
chlorides	1	-2.002275	0.398076	-5.03	<.0001
total_sulfur_dioxide	1	-0.002372	0.000506	-4.68	<.0001
pH	1	-0.435192	0.116037	-3.75	0.0002
sulphates	1	0.888668	0.110042	8.08	<.0001
alcohol	1	0.290673	0.016811	17.29	<.0001

Model: MODEL1
Dependent Variable: quality



Model: MODEL1
Dependent Variable: quality





Conclusion for Red Wine Quality

Starting with 11 dependent variables that might theoretically be good predictors of wine quality, a forward section technique regression model was used to reduce them to 6 variables: alcohol, volatile acidity, sulfates, total_sulfur_dioxode, chlorides, and pH. Then the result of the graphs shows that the winequalityRed dataset doesn't support the linear regression model because the dataset is not independent of the data point and not fits a diagonal line. Additionally, all residual plots have patterns, meaning the model's not picking up all the signals of the dataset. Might add to a variable that helps to improve model results or apply a nonlinear model.

Multilinear Regression Model to the "winequalityWhite" dataset in SAS Studio.

I wanted to predict which independent variables were to affect the quality of wine; what would be a great wine to make that prediction? I did the same step as in figure 1 with the "winequalityWhite" dataset in SAS Studio. Let's start the ANOVA table, the number of observations, and read 4898, which means the dataset doesn't have a null or miss variable. P-value is very small, 0.001, which tells us the model is doing very well and has much variability in my target. F value helps to know how a group of variables is jointly significant as the result of F-value 174.34 is high, which means the group of the joint is not good. SS also tells how the model fits the data set; the best number is zero as our model SS value's 1082.66. The result of RMSE is 0.7 is a low number. The R-square (coefficient of determination) value shows the proportion of the variability of my input and how strong the relationship between the two variables is. R-square values between 0 to 1 and higher values are better as our result is 0.28 as 28%, so low to apply those variables. The following table shows Parameter Estimate has the most critical piece of information because the P- value showed the model is fit data, but F-value and R-square gave the opposite. Other result shows that the independent variable is not significant relation each other. As we can see, for each continuous variable's p-value, some are high, some are lower than 0.05, and the high value can badly affect the R square and F-value. Additionally, the table has Standard Error helps to know how sample means are broadly spread around the population mean. The best result to close zero tells the sample is representative of the population as our result "density" has a high number of 19.07 which can affect the model result and quality of the wine.

Thus, when eliminating higher p-value from the parameter estimates, the linear model should have fixed_acidity, volatile_acidity, residual_sugar, free_sulfur_dioxide, density, pH, sulphates, and alcohol.

Figure 6

The result of the ANOVA table of winequalityWhite in SAS Studio.

9/8/22, 12:06 PM Results: Program 1

Model: MODEL1
Dependent Variable: quality

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	11	1082.66119	98.42374	174.34	<.0001
Error	4886	2758.32860	0.56454		
Corrected Total	4897	3840.98979			

Root MSE	0.75136	R-Square	0.2819
Dependent Mean	5.87791	Adj R-Sq	0.2803
Coeff Var	12.78272		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	150.19283	18.80418	7.99	<.0001
fixed_acidity	1	0.06552	0.02087	3.14	0.0017
volatile_acidity	1	-1.86318	0.11379	-16.37	<.0001
citric_acid	1	0.02209	0.09577	0.23	0.8176
residual_sugar	1	0.08148	0.00753	10.82	<.0001
chlorides	1	-0.24728	0.54654	-0.45	0.6510
free_sulfur_dioxide	1	0.00373	0.00084415	4.42	<.0001
total_sulfur_dioxide	1	-0.00028575	0.00037806	-0.76	0.4498
density	1	-150.28417	19.07451	-7.88	<.0001
pH	1	0.68634	0.10538	6.51	<.0001
sulphates	1	0.63148	0.10039	6.29	<.0001
alcohol	1	0.19348	0.02422	7.99	<.0001

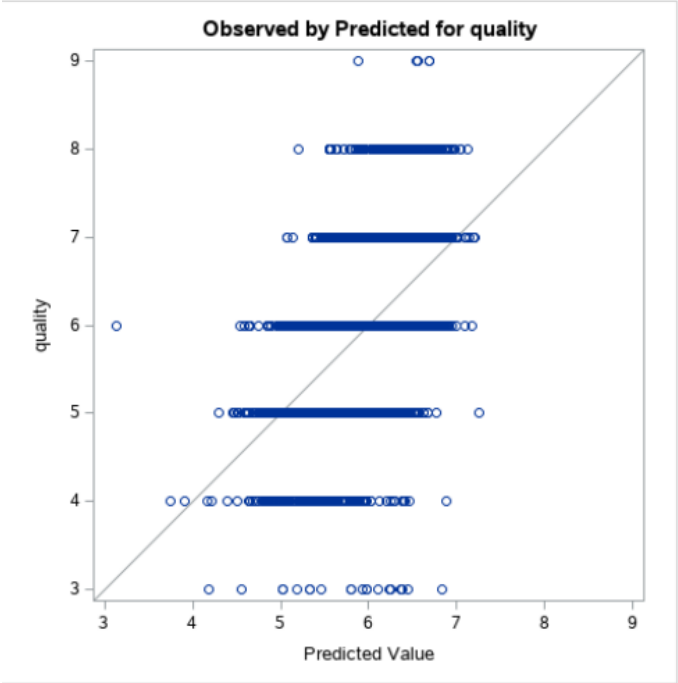
The following plot charts show more details about the white wine quality dataset in the SAS Studio dataset. The Observed by Predicted for quality indicates that a data point stays around the diagonal

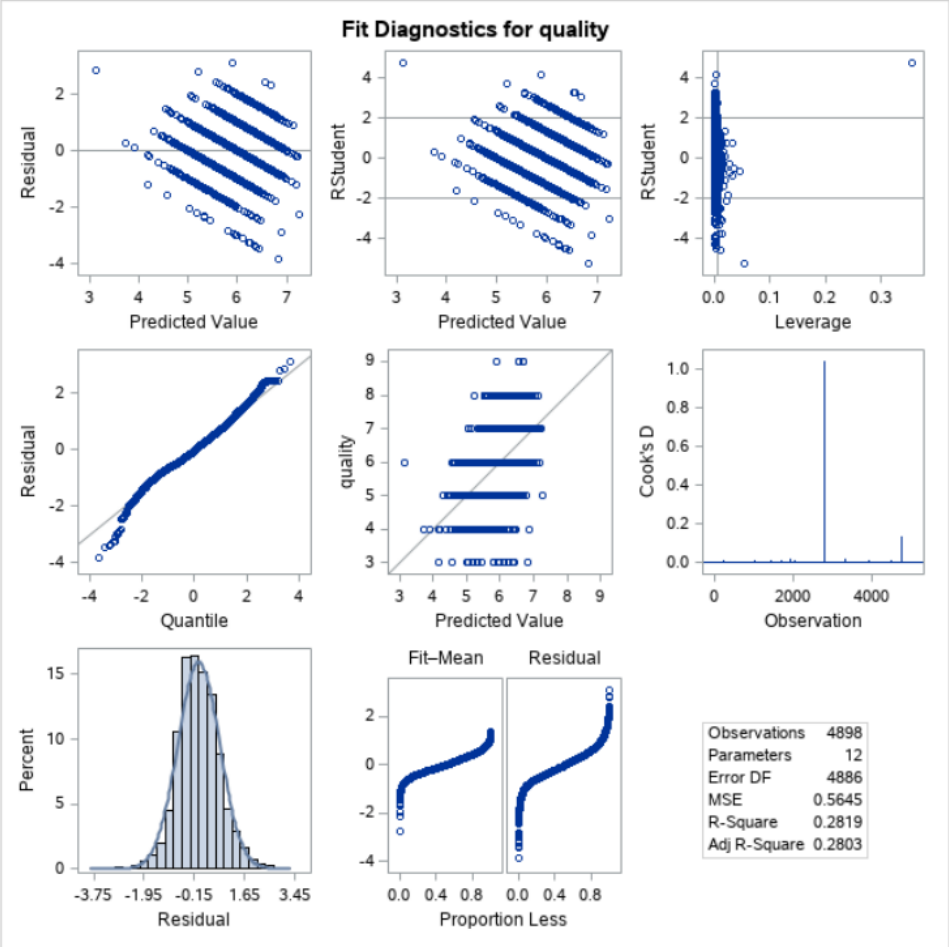
line; the best result points should be close to the line and randomly scatter, not a pattern. Our result doesn't have a random scatter, meaning one or more of our inputs are insufficient to help us, or we should look at a different model. The Fit Diagnostics for quality allows us to validate our assumption to meet to perform a linear regression. First is the linear relationship between the dependent variable and the input variables. Q-Q shows a linear relationship between residuals and the target variable. Another assumption is the independence of the individual observation, as the first chart shows that the plot is not independent, which means the plot has a pattern. The Residual, RStudent, and Quality plots show no random scatter, meaning the dataset does not fit the model. We can see one point, and the other point has a dependence. The third assumption must be Normal distribution; the bottom of the first column chart shows a skew distribution. Q-Q and Cook's D show dataset has outliers, and Cook's D has one highest outlier point that can affect the model.

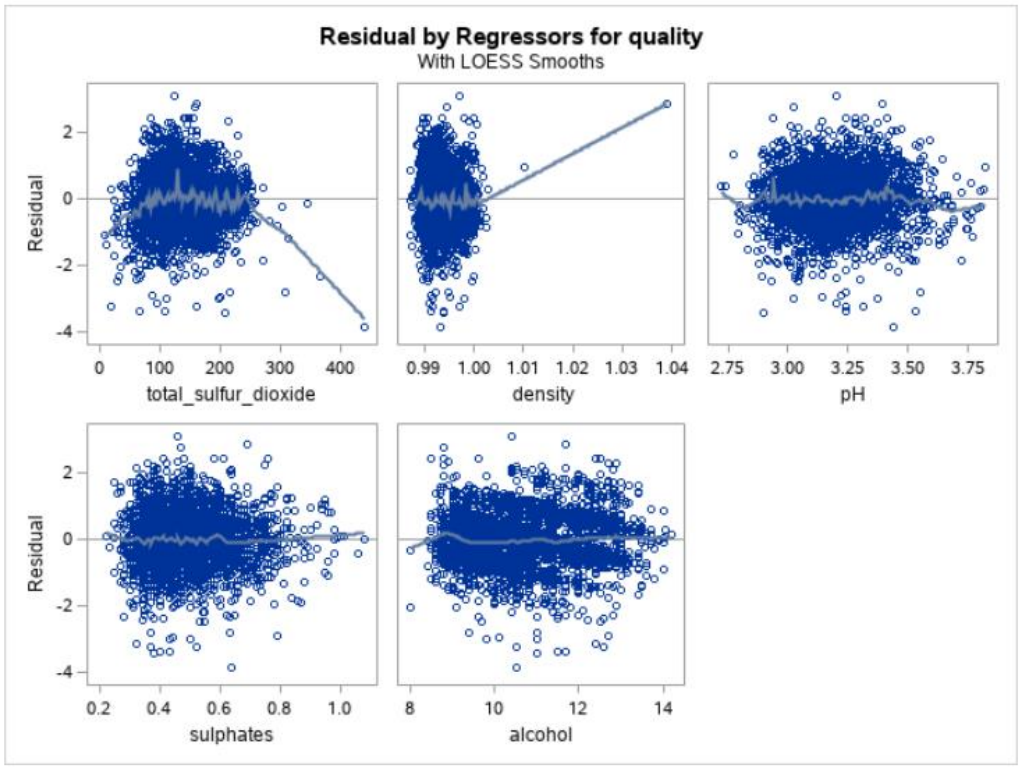
The last one is an equal error, meaning no plot patterns on the plot. As shown in The Residual by Regressors for quality, LOESS helps to see the pattern because data is so noisy that it is so hard to see any pattern without LOESS. As a result, all plot has shaped pattern that a linear model is not fit for the winequalityWhite dataset.

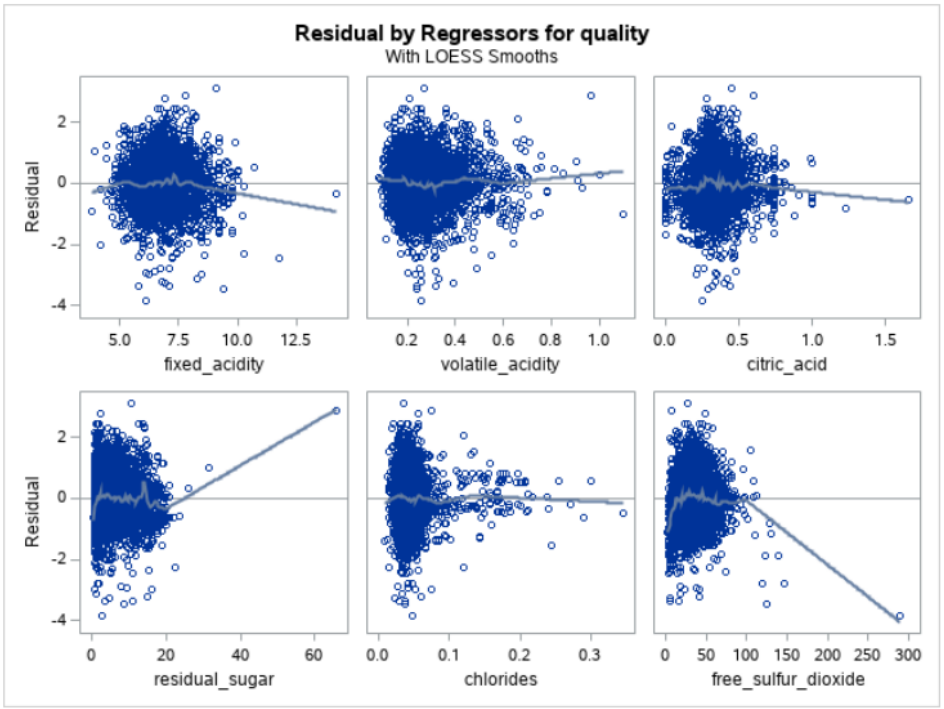
Figure 7

The result of the multilinear regression for the "winequaliteWhite" dataset in SAS Studio.


Observed by Predicted for quality
quality
Predicted Value


Fit Diagnostics for quality
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
quality
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 4898
Parameters 12
Error DF 4886
MSE 0.5645
R-Square 0.2819
Adj R-Square 0.2803


Residual by Regressors for quality
With LOESS Smooths
Residual
total_sulfur_dioxide
density
pH
sulphates
alcohol


Residual by Regressors for quality
With LOESS Smooths
Residual
fixed_acidity
volatile_acidity
citric_acid
residual_sugar
chlorides
free_sulfur_dioxide

Forward Selection technique for building multilinear regression model by "winequalityWhite" into SAS Studio

The Forward Section technique is unique and valuable for filtering which independent variable can predict the target value. I used the SAS Studio, and each step is the same in figure 4 with the winequalityWhite dataset. And I added LOESS to eliminate the noisy data part to see precise movement in data points. Figure 8 has model and dataset information about the "winequalityWhite" dataset. The number of observations reads and used, 4898, means the dataset doesn't have missing data. The Forward Selection Summary display eight variable in the model and sort by SBC result from most to less alcohol, volatile_acidity, residual_sugar, free_sulfur_dioxide, density, pH, sulphates, and fixed_acidity. The Forward selection technique filter the total_sulfur_dioxide might increase the variable's SBC result. The Fit Criteria for quality has AIC, AICC, SBC, and Adj R-Sq information. Those help to compare models less effect is considered the best. As the line chart shows, AIC, AICC, and SBC have the same result chart for continuous variables.

Figure 8

The Forward Section technique in multiLinear regression model's step of "winequalityWhite" dataset in SAS Studio.

Data Set	WORK.WINEQUALITYWHITE
Dependent Variable	quality
Selection Method	Forward
Select Criterion	SBC
Stop Criterion	SBC
Effect Hierarchy Enforced	None

Number of Observations Read	4898
Number of Observations Used	4898

Dimensions	
Number of Effects	12
Number of Parameters	12

Forward Selection Summary

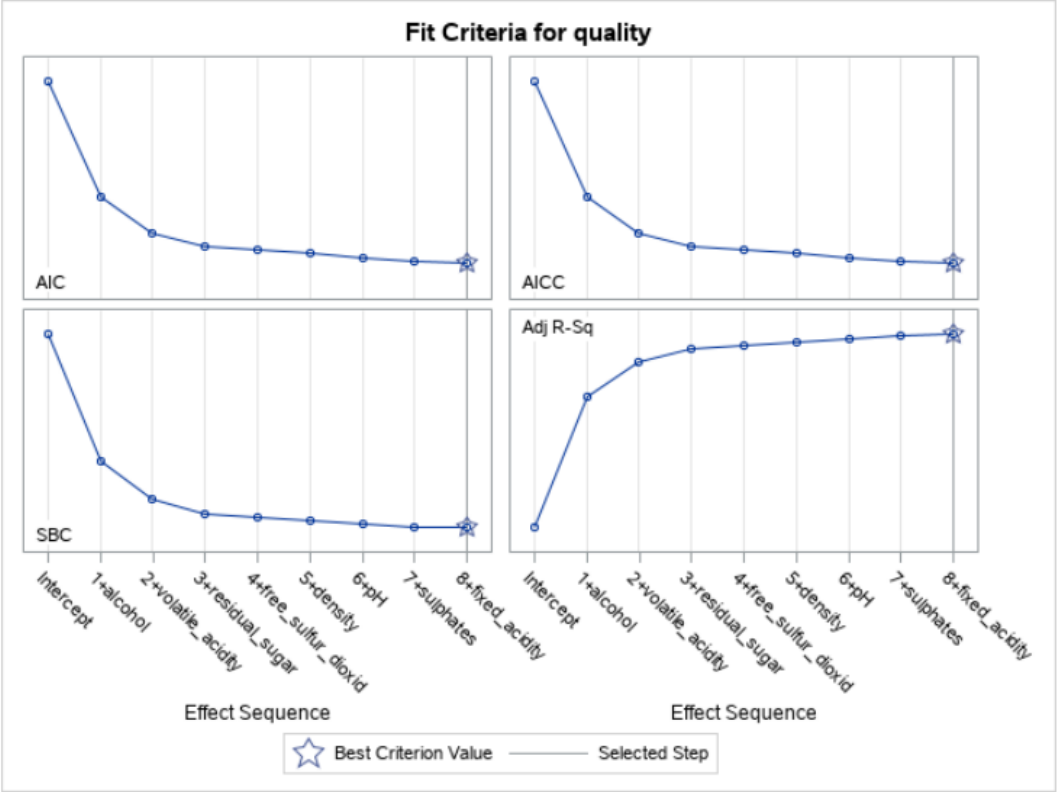
Step	Effect Entered	Number Effects In	SBC
0	Intercept	1	-1182.1919
1	alcohol	2	-2204.1463
2	volatile_acidity	3	-2510.8805
3	residual_sugar	4	-2621.7693
4	free_sulfur_dioxide	5	-2649.5273
5	density	6	-2674.1327
6	pH	7	-2707.5595
7	sulphates	8	-2732.5423
8	fixed_acidity	9	-2735.1649*

* Optimal Value of Criterion

Selection stopped at a local minimum of the SBC criterion.

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

Let's move statistical information in the ANOVA table to get each independent behavior in the model and its relation to the target value. First, the Analysis of Variance table has SS, MS, F value, and p-value. The p-value shows the best result of 0.001 means the dataset strongly supports the model as statistically significant. F value and SS result are high 239.73 and 1082.20 both marks close the zero for the best; F value tells the group of variables doesn't jointly well. SS result shows that the model doesn't fit the dataset. RMSE result 0.75 and R square 0.28, a low number for r square, means independent and target variables have a weak correlation.

The Parameter Estimates table has a t-value and p-value for each variable. The model uses a stopping rule to filter lower than 0.05 p-value; eight variables are statistically significant. The

standard error for “density” is a high number that might affect more to the target variable. The t value is used to compare the mean of the sample tests.

Thus, the linear regression formula might have all eight variables; fixed_acidity, volatile_acidity, residual_sugar, free_sulfur_dioxide, density, pH, sulphates, and alcohol.

The plot of the Observed by Predicted for quality shows data points have patterns and are not close to the diagonal line, which means our inputs are insufficient, as we should look at a different model. The following graph is Fit Diagnostics for quality help to see how data points stay around the line as Residual and RStudent plots have a pattern that means the dataset doesn’t have independence (random scatter). Q-Q has a linear fit and also shows outliers. The Cook’s D has one point pretty high that should be an essential outlier to affect model results. Residual shows skew distribution. The Residual-fit Spreat plot shows that the right residual is taller than the left, which means the model can’t explain all the variations.

The final graph has a data point detail as I added LOESS to see data point movement that shows all variables have a pattern and tells us the dataset does not support the linear regression assumption.

Figure 9

The result of the Forward Section technique for the “winequaliteWhite” dataset in SAS Studio.

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

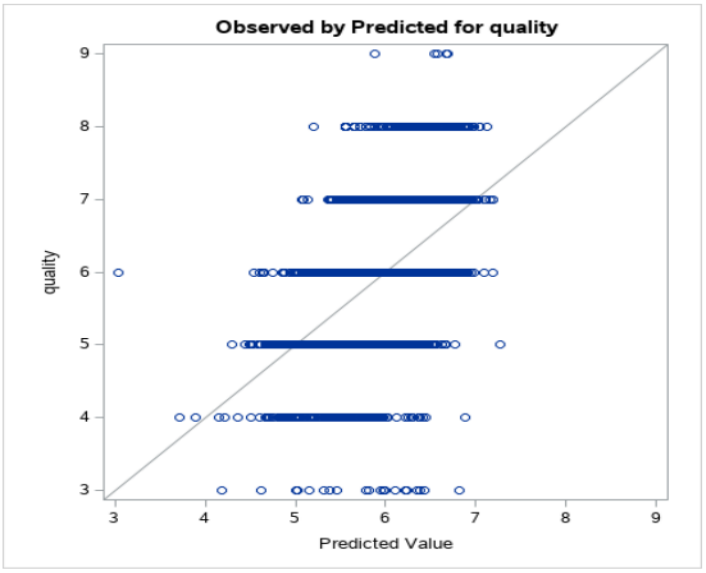
Note: The p-values for parameters and effects are not adjusted for the fact that the terms in the model have been selected and so are generally liberal.

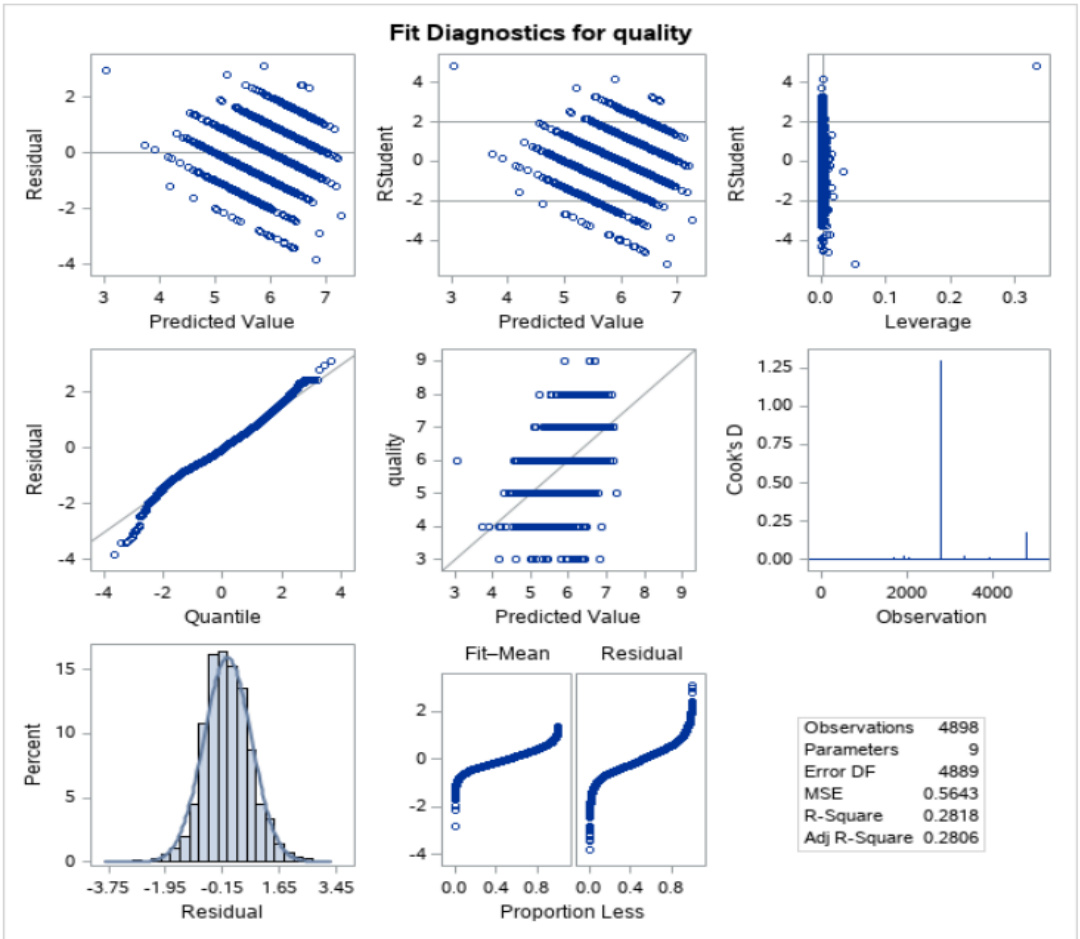
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	1082.20642	135.27580	239.73	<.0001
Error	4889	2758.78338	0.56428		
Corrected Total	4897	3840.98979			

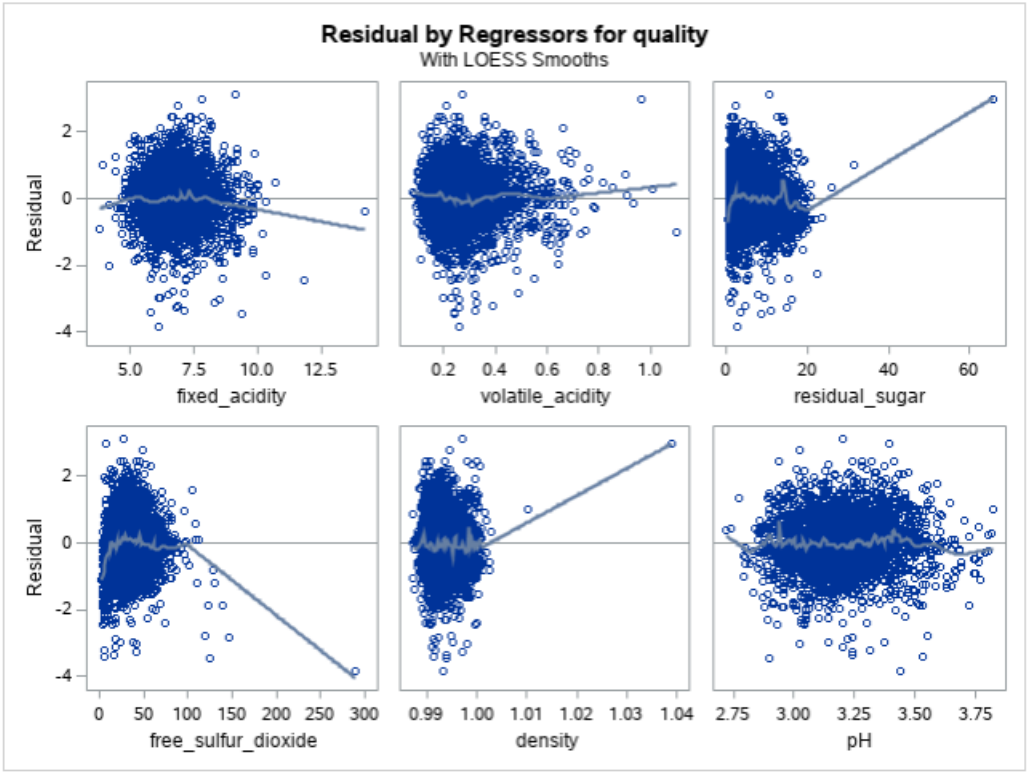
Root MSE	0.75119
Dependent Mean	5.87791
R-Square	0.2818
Adj R-Sq	0.2806
AIC	2106.36588
AICC	2106.41090
SBC	-2735.16488

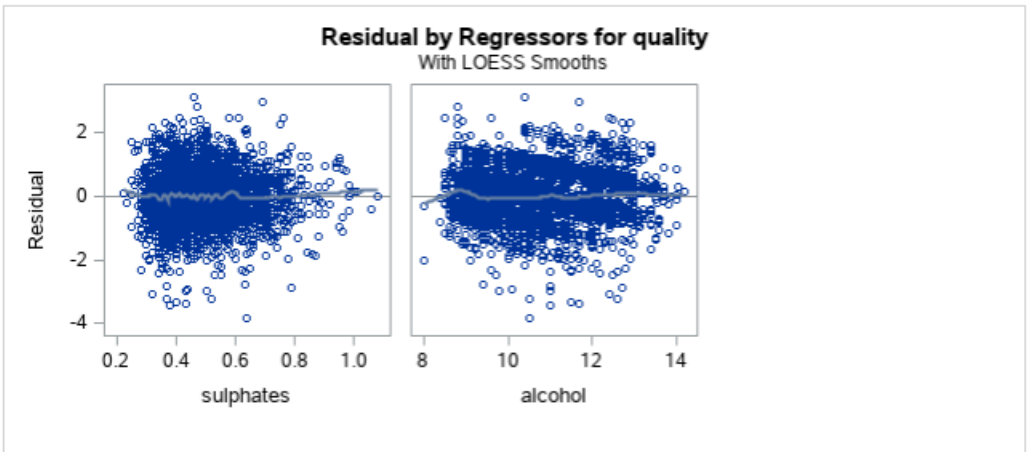
Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	154.106248	18.100130	8.51	<.0001
fixed_acidity	1	0.068104	0.020431	3.33	0.0009
volatile_acidity	1	-1.888140	0.109509	-17.24	<.0001
residual_sugar	1	0.082847	0.007287	11.37	<.0001
free_sulfur_dioxide	1	0.003349	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

Model: MODEL1
Dependent Variable: quality









Conclusion for White Wine Quality

Starting with 11 dependent variables that might theoretically be good predictors of wine quality, a forward section technique regression model was used to reduce them to 8 variables: fixed acidity, sulfates, ph, free sulfur dioxide, residual square, volatile acidity, and alcohol. Then the result of

the graphs shows that the winequalityWhite dataset doesn't support the linear regression model because the dataset is not independent of the data point and not fits a diagonal line. Additionally, all residual plots have patterns, meaning the model's not picking up all the signals of the dataset. Might add to a variable that helps to improve model results or apply a nonlinear model.

Reference

UCI Machine Learning Repository: Wine Quality Data Set

About Linear Regression | IBM