Transactions on Economics, Business and Management Research
ISSN: 2960-1789, eISSN: 2960-2254 | Volume 10
MEEA 2024

The Research on Influence Factors of Red Wine Quality

Tianze Liu*

College of China University of Geosciences, Beijing, 100083, China

* Corresponding Author Email: 1019211119@email.cugb.edu.cn

Abstract. Nowadays, product quality certificates are being used by enterprises to market their products. This is an expensive and ineffective process that takes a long time and requires assessments from customers and human experts. Determining the relationship between a red wine's chemical composition and subjective quality is a difficult task. This paper investigates the use of statistical methods, such as linear regression, to predict the values of target variables and ascertain how dependent the target variable is on the independent factors. It is more accurate than previous methods. These results contribute to the understanding of how different red wine consumers perceive quality and can help the red wine industry identify the primary sensory-active ingredients that influence quality in different red wines. The paper uses linear regression to solve the problem. The research examines at those factors' VIF value and significance in order to assess the efficacy of this operation. It turns out that volatile acidity, chlorides, total sulfur dioxide, sulphates and alcohol have a significant linear relationship with red wine quality, while the other factors have less influence on red wine quality. Overall, red wine quality can be evaluated based on how much these variables affect them.

Keywords: Linear regression; Influence factors; Red wine quality.

1. Introduction

As an essential component of the lives of the general population, red wine quality has always been a substantial issue for people around the world. Wine polyphenols may be useful in reducing the risk of cardiovascular disease (CVD). Clinical research indicates that moderate wine drinking may have an impact on CVD [1]. Red wine quality can be influenced by many factors. Quality perception among experts and consumers is associated with chemical composition [2]. Significant relationships were discovered in 1974 between (a) total anthocyanins, colored and non-colored anthocyanins, pH, flavor and color scores, and non-colored anthocyanins; (b) total anthocyanins and pH; (c) aroma and total pigments; (d) the chemical parameters of pH, total pigments, and total anthocyanins. In 1975, pH and non-colored anthocyanins were also associated with flavor [3]. However, studies on the chemical components involved in quality perception are limited. A critical gap exists in the understanding of the factors influencing red wine quality. So, it is critical to comprehend the elements that affect the quality of red wine.. Therefore, this paper intends to assist people in better understanding the relationship between chemical components and red wine quality based on the dataset.

Red wine is a intricate substance with numerous factors involved in red wine quality. Predictions of red wine quality is also a significant for red wine industries and customers. Focusing on relatively simple mixes of fragrance compounds and tastants comprising sensory-active molecules at supra-, peri-, and sub-threshold concentration levels, psychophysicists research the sensory properties (quantitative and qualitative aspects) [4, 5]. Legin and his team evaluate Italian wine by the electronic tongue (recognition, quantitive analysis and correlation with human sensory) [6]. However, only twenty samples of Barbera d'Asti and thirty-six samples of Gutturnio wine were utilized for the measurements, which is limited in Italian wine and relatively small. This paper use larger database to get a more common conclusion. Cortez used data mining to access the data, which is relatively large [7]. In order to assess the models' accuracy, he employed multiple linear regression models for analysis and estimated the parameters using Ordinary Least Squares (OLS) regression. However, he

did not consider the interaction between each factor. Ferrer applied a quality loss function to represent wine quality reduction [8]. But that model did not take many factors into account, such as density, sulfur dioxide and residual sugar. Thus the method still need improvement. Although some academics have evaluated wine quality using machine learning approaches, there is still plenty of potential for improvement. Sun et al. used neural networks fed with fifteen input factors to predict six regional origins of wine [9]. For their experiments, 170 data samples from Germany were utilized. They got a predictive rate nearly 100%. But the problem is still that the database is relatively small, resulting the result is not accurate. Neural networks were also utilized by Vlassides et al. to classify Californian wine [10]. Wine classification is based on chemical analysis and grape maturity level. Experiments with a sample of thirty-six cases yielded an error rate of only six percent. But the method still needs to be improved. A more accurate prediction can be achieved by taking into account a subset of features rather than all of the variables.

This paper focuses on variables (fixed acidity, volatile, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, PH, sulphates, alcohol) that were studied to determine their impact on red wine quality, and choose an appropriate model to examine the relationship between these variables and the quality of red wine.

In conclusion, the multiple linear regression model will be utilized in this article to examine how these 11 characteristics affect the quality of red wine. By choosing the essential wine characteristics that are crucial in defining the quality of the wine, it is possible to create a model with an interface that forecasts the wine's quality.

2. Methods

2.1. Data Source

The data is taken from the Kaggle website. This data contains some characteristics of red wine, with a total of 1600 observations. They were all considered as samples for this study. The original dataset remained in .csv format [7].

2.2. Variable Selection

There are a lot of null values for variables in the original dataset, which contains a huge quantity of data. Only the physicochemical and sensory variables are available due to logistical and privacy concerns. Information such as grape varietals, wine brands, and wine prices is lacking. The data contains 12 variables (fixed acidity, volatile, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, PH, sulphates, alcohol) and one dependent variable (red wine quality). Table 1 displays a detailed description of this dataset.

Table 1. List of variables

variable	Logogram	meaning
fixed acidity	x_1	the fixed acidity of red wine
volatile	x_2	the volatility of red wine
citric acid	x_3	the number of citric. Acid
residual sugar	x_4	the number of residual sugar
chlorides	x_5	the number of chlorides
free sulfur dioxide	x_6	the number of free sulfur dioxide
total sulfur dioxide	x_7	the number of total sulfur dioxide
density	x_8	the density of red wine
PH	x_9	the PH's number
sulphates	x_{10}	the sulphates' number
alcohol	x_{11}	alcohol content of red wine
quality	Y	the quality of red wine

2.3. Method Introduction

One of the most widely used statistical methods for linking a group of two or more variables is the multiple linear regression model. It serves as a basis for the linear relationship that exists between the variable being explained and several additional explanatory factors. It serves as a basis for the linear relationship that exists between the variable being explained and several additional explanatory factors. Furthermore, estimating a set of parameters is its fundamental concept in order to minimize the sum of squares of the residuals between the independent and dependent variables.

3. Results and Discussion

3.1. Simple Linear Regression

This paper's examination demonstrates the wide range of elements that affect the quality of red wine. The paper analyses the relationship between single variable and red wine quality. Take fixed acidity, volatile acidity and residual sugar as example. The typical linear regression mathematical model is:

$$E(Y) = \beta_0 + \beta_1 x \tag{1}$$

In the formula above: β_0 is a constant term. The result is shown below:

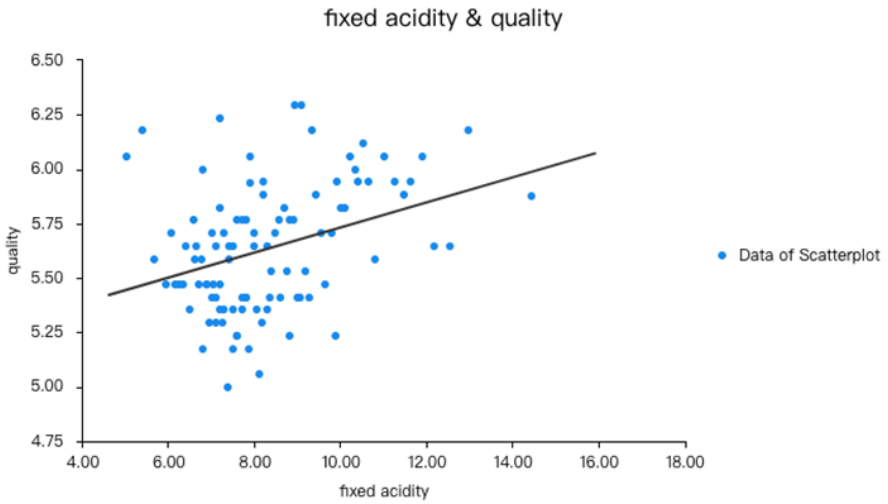

Figure 1. Linear Regression Between Fixed Acidity and Quality

The result is: y = 5.157 + 0.058*fixed acidity. From Figure 1, it can be seen fixed acidity has a positive effect on quality.

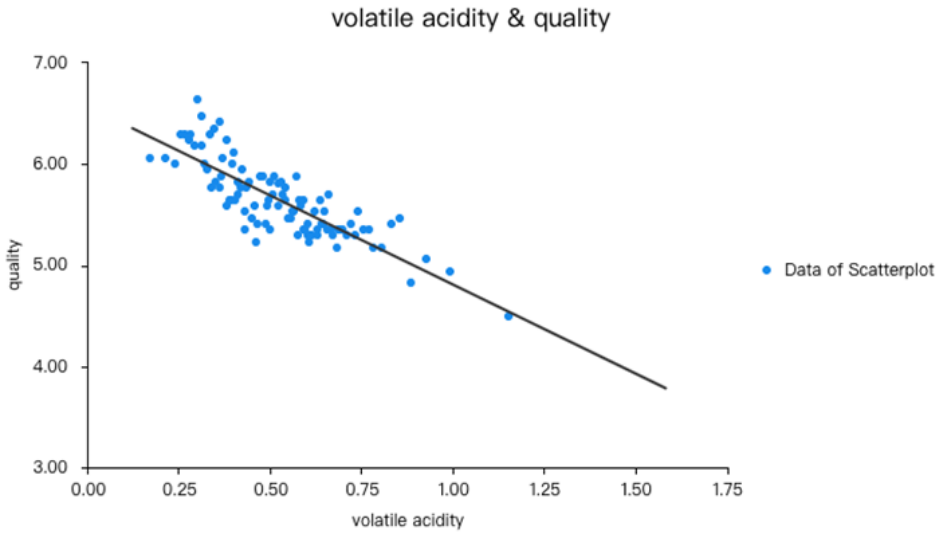

Figure 2. Linear Regression Between Volatile and Quality

The result is: quality = 6.566-1.761*volatile acidity. From Figure 2, it can be seen volatile acidity has a negative effect on quality. The coefficient of the formula is -1.761, which demonstrates the result.

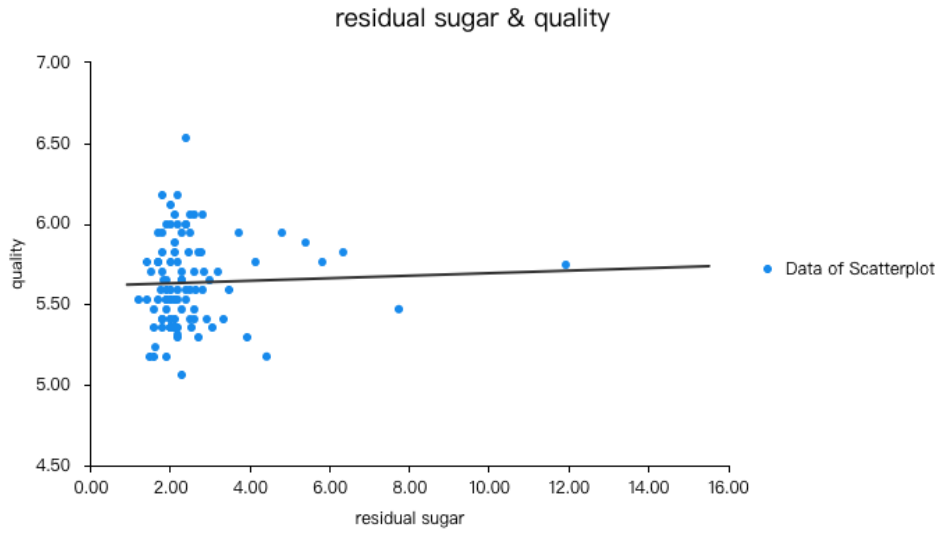

Figure 3. Linear Regression Between Residual Sugar and Quality

The result is: quality = 5.616 + 0.008*residual sugar. From Figure 3, it can be seen residual sugar nearly does not have effect on quality. Similarly, it can be calculated that free sulfur dioxide, sulphates and alcohol have positive effect on red wine quality. While volatile acidity, chlorides, total sulfur dioxide and pH have negative effect on red wine quality. The other variables do not have effect on red wine quality.

3.2. Correlation Results

After considering the relationship between single variable and red wine quality, the relationship between all these factors and red wine quality are needed to be considered. Table 2 demonstrates the Pearson Correlation between each variable and red wine quality.

Table 2. Correlation results Between Dependent and Independent Variables

variable	Pearson Correlation
alcohol	0.476**
sulphates	0.251**
pH	-0.058*
density	-0.175**
free sulfur dioxide	-0.051*
citric acid	0.226**
volatile acidity	-0.391**
total sulfur dioxide	-0.185**
fixed acidity	0.124**
chlorides	-0.129**
residual sugar	0.014

* p<0.05 ** p<0.01

From Table 2, the Pearson correlation coefficient between these factors and the quality of red wine is presented. The research data found that alcohol, sulphates and volatile are, in turn, the variables that have greatest positive correlations with red wine quality. While other variables, such as citric acid, residual sugar, density and fixed acidity, do not contribute much to the quality of red wine.

3.3. Multiple Linear Regression

The general mathematical model for multiple linear regression is:

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{12} x_{11} + e \quad (2)$$

In the formula above: β_0 is a constant term, and e is a residual term.

Table 3. Linear Regression coefficients

	B	S.E.	Beta	T	significance	VIF	Tolerance
Constant	21.965	21.195	-	1.036	0.300	-	-
volatile acidity	-1.084	0.121	-0.240	-8.948	0.000**	1.789	0.559
citric acid	-0.183	0.147	-0.044	-1.240	0.215	3.128	0.320
residual sugar	0.016	0.015	0.029	1.089	0.276	1.703	0.587
chlorides	-1.874	0.419	-0.109	-4.470	0.000**	1.482	0.675
free sulfur dioxide	0.004	0.002	0.056	2.009	0.045*	1.963	0.509
total sulfur dioxide	-0.003	0.001	-0.133	-4.480	0.000**	2.187	0.457
density	-17.881	21.633	-0.042	-0.827	0.409	6.344	0.158
sulphates	0.916	0.114	0.192	8.014	0.000**	1.429	0.700
alcohol	0.276	0.026	0.364	10.429	0.000**	3.031	0.330
pH	-0.414	0.192	-0.079	-2.159	0.031*	3.330	0.300
fixed acidity	0.025	0.026	0.054	0.963	0.336	7.768	0.129

The multiple linear regression equation model's regression coefficients are displayed in Table 3. The p-values of the T-test for volatile acidity, chlorides, total sulfur dioxide, sulphates and alcohol did not exceed 0.003. Therefore, it can be considered that all five independent variables have a considerable

influence on the dependent variable(quality). The following multiple linear regression equation can be calculated based on the given data:

$$E(Y) = 21.965 - 1.084x_1 - 0.183x_2 + \cdots + 0.025x_{11} \quad (3)$$

The multivariate correlation coefficient R obtained from this model definition is 0.600, the coefficient R-squared for fitting multiple linear regression is 0.361, and the adjusted R-squared is 0.356. The RMSE is 0.646. Thus, the model has a good fit.

4. Conclusion

The study selected 1600 samples from the Kaggle website, which includes 12 variables. The method (Multiple linear regression analysis) is precise, efficient, and all-encompassing. Because it obtains the Pearson correlation coefficients for each variable after conducting a multifactor analysis.

In order to determine whether there may be a relationship between the variables and the quality of red wine, the article employs a multiple linear regression model throughout the analysis stage. In order to get more precise results, the study includes interaction terms with coefficients in the equation and accounts for interaction effects. Therefore, the factors that positively impact on red wine quality are sulphates, alcohol, free sulfur dioxide. The volatile acidity, pH, total sulfur dioxide, and chlorides are negatively correlated with red wine quality. From all of variables, citric acid, residual sugar, density, fixed acidity are not the main factors. They have little influence on red wine quality.

With the model, customers longing for better red wine quality have a preference for sulphates, alcohol, free sulfur dioxide. However, there are still some deficiencies such as the data is not the most recent version, the sample size is limited, and no correlations between the variables can be established. Meanwhile, the data does not cover all kinds of red wine and all the factors of red wine. To make this improvement, look for new data and investigate potential causal relationships between variables and red wine quality using the control variable approach.

References

[1] L. Snopek, et al. Contribution of red wine consumption to human health protection, Molecules. 23(7) (2018) 1684.

[2] M.P. Sáenz-Navajas, et al. Sensory-active compounds influencing wine experts' and consumers' perception of red wine intrinsic quality. LWT-Food Science and Technology. 60(1) (2015) 400-411.

[3] M.G. Jackson, et al. Red wine quality: Correlations between colour, aroma and flavour and pigment and other parameters of young Beaujolais, Journal of the Science of Food and Agriculture. 29(8) (1978) 715-727.

[4] P. Breslin, Human gustation and flavour. Flavour and Fragrance Journal, 16(6) (2001) 439-456.

[5] A. Legin, et al. Evaluation of Italian wine by the electronic tongue: recognition, quantitative analysis and correlation with human sensory perception, Analytica Chimica Acta. 2003, 484(1) (2003) 33-44.

[6] P. Cortez, et al. Modeling wine preferences by data mining from physicochemical properties, Decision support systems. 47(4) (2009) 547-553.

[7] J.C. Ferrer, et al. An optimization approach for scheduling wine grape harvest operations. International Journal of Production Economics. 112(2) (2008) 985-999.

[8] L.X. Sun, et al. Classification of wine samples by means of artificial neural networks and discrimination analytical methods, Fresenius' journal of analytical chemistry. 359 (1997) 143-149.

[9] S. Vlassides, et al. Using historical data for bioprocess optimization: modeling wine characteristics using artificial neural networks and archived process information, Biotechnology and Bioengineering. 73(1) (2001) 55-68.

[10] Y. Gupta, Selection of important features and predicting wine quality using machine learning techniques, Procedia Computer Science. 125 (2018) 305-312.