Assess the Quality of a Bottle of Wine: Approaches from a Regression Model

Author:

Ziyu Wang

Xi Liu

Salimata Diakite

Abstract

Wine is a kind of well-loved alcoholic beverage all around the world. People like to use the quality of wines to assess their value. However, judging the value of wine is very subjective as it involves the personal preferences of the tasters.

The article is aimed at Providing an objective and systematic wine scoring mechanism by establishing a model to find out the influence of the objective factors of wines (e.g. PH values) on the expert-graded wine quality.

In the model, we first used a series of predictive variables to predict the response variables via OLS regression method, then we performed a regression diagnosis of the model to improve the accuracy of fitting. At last we select the “best” regression model by using the Akaike Information Criterion.

problem analysis

According to the data of red wine and white wine we have, we see that there are eleven factors may influence the wine quality. They are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol. However, we do not know how much they influence the wine quality respectively and whether each of them is useful to the model. Thus, our goal is to select the valid parameters, acquire their coefficients and reject the invalid parameters.

Model solving

First we built a rough polynomial regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{12} X_{12}$$

For which,

Y	The expert-graded wine quality between 0 (very bad) and 10 (very excellent)
$X_i\ (i = 1, \cdots, 12)$	The numerical value of each objective factor
β_0	intercept term
$\beta_i\ (i = 1, \cdots, 12)$	The regression coefficient of each objective factor

(table 1)

We used R studio to acquire two simple OLS regression fitting models of red wine and white wine:

$$Y_{RED} = 21.965 + 0.025X_1 - 1.084X_2 - 0.183X_3 + 0.016X_4 - 1.874X_5 + 0.004X_6 - 0.003X_7 - 17.881X_8 - 0.414X_9 - 0.916X_{10} + 0.276X_{11}$$

$$Y_{WHITE} = 150.193 + 0.066X_1 - 1.084X_2 - 1.863X_3 + 0.022X_4 + 0.081X_5 + 0.004X_6 - 0.0003X_7 - 150.284X_8 + 0.686X_9 + 0.631X_{10} + 0.193X_{11}$$

But these two regression fitting models are very simple and rough. There is no conclusion as to whether the model is appropriate, and our confidence of the model parameters, in most cases, depends on the extent to which they meet the statistical assumptions of the OLS model.

In the mean time, we found that there are two factors called free sulfur dioxide and total sulfur dioxide among the eleven factors, we wondered if there is an interaction between these two.

We use R to test and have found that

free.sulfur.dioxide:total.sulfur.dioxide	P value	Exist an interaction or not
Red wine	0.3180	FALSE
White wine	< 2e-16	TRUE

(table 2)

We noticed that there is no significant interaction between free.sulfur.dioxide and total.sulfur.dioxide in the red wine quality test, but we acquire exactly the reverse result in the white wine quality test so we rejected the free.sulfur.dioxide:total.sulfur.dioxide term in the red wine quality test and kept it in the white wine quality test.

We also need to check whether the models have multicollinearity problem, since The least Square estimates are not reliable when there exists multicollinearity in adjustment model. We can use statistical magnitude VIF(Variance Inflation Factor) to test whether the models exist multicollinearity problem. If $\sqrt{vif} > 2$, it suggests that the model exists multicollinearity problem.

We use R studio to conduct the test and the results are as follows:

Red wine		White wine	
fixed.acidity	TRUE	fixed.acidity	FALSE
volatile.acidity	FALSE	volatile.acidity	FALSE
citric.acid	FALSE	citric.acid	FALSE
residual.sugar	FALSE	residual.sugar	TRUE
chlorides	FALSE	chlorides	FALSE
free.sulfur.dioxide	FALSE	free.sulfur.dioxide	FALSE
total.sulfur.dioxide	FALSE	total.sulfur.dioxide	TRUE
density	TRUE	density	TRUE
pH	FALSE	pH	FALSE
sulphates	FALSE	sulphates	FALSE
alcohol	FALSE	alcohol	FALSE
		free.sulfur.dioxide:total.sulfur.dioxide	TRUE

(table 3)

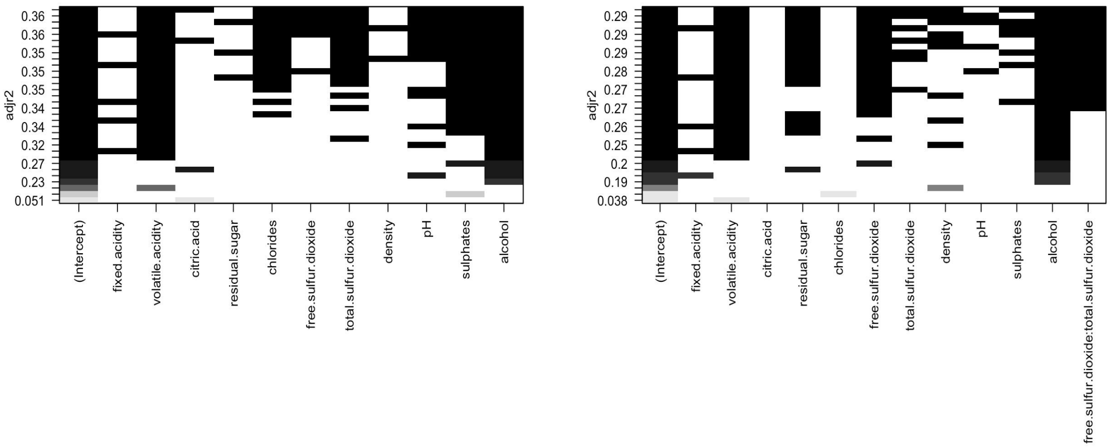
We can see from the above table that both red wine and white wine exist multicollinearity problem.

After having preliminarily analyzed the model, we realized that not all variables have significant influence on the expert-graded wine quality, so we] need to improve the model fitting degree and simplicity by erasing some redundant variables.

We chose to use the AIC (Akaike Information Criterion) value to compare the models that contain different variables, as it estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.

We have two methods of selecting the optimal variables from a large number of predictive variables, stepwise method and all-subsets regression method. In most cases, all-subsets regression method is better than stepwise method. However, good imitative effect, sometimes, is not as helpful as we thought to the model, so we choose all-subsets regression method as a aiding method and use stepwise method to determine the final model.

First we used R studio to conduct all-subsets regression method, we created two charts to make sure the results are more intuitive:



Red wine

White wine

(chart 1)

In the two charts, the larger the ordinate value (R^2) is, the higher the fitting degree will be. The black areas represent the variables need to be kept of the given (R^2), while the white areas represent they can be rejected.

Next we used R studio to conduct backward stepwise method, the results are as follow:

	Red wine		White wine	
Intercept	4.441160		144.5	
Variables	Kept or not	coefficients	Kept or not	coefficients
fixed.acidity	FALSE	\	FALSE	0.06934
volatile.acidity	TRUE	-1.011272	TRUE	-1.761
citric.acid	FALSE	\	FALSE	\
residual.sugar	FALSE	\	TRUE	0.07767
chlorides	TRUE	-2.022234	TRUE	\
free.sulfur.dioxide	TRUE	0.005148	TRUE	0.02078
total.sulfur.dioxide	TRUE	-0.0033500	TRUE	0.002765
density	FALSE	\	TRUE	-145.1
pH	TRUE	-0.485986	TRUE	0.6689
sulphates	TRUE	0.882581	TRUE	0.6201
alcohol	TRUE	0.289259	TRUE	0.197
free.sulfur.dioxide: total.sulfur.dioxide	\	\	TRUE	-0.0001

(table 4)

Comparing the table 4 with chart 1, the selections of the optimal variables are very close, so we use the data in table 4 to generate two fitting models of red wine and white wine.

Therefore, the relation between objective factors and expert-graded wine quality can be written as:

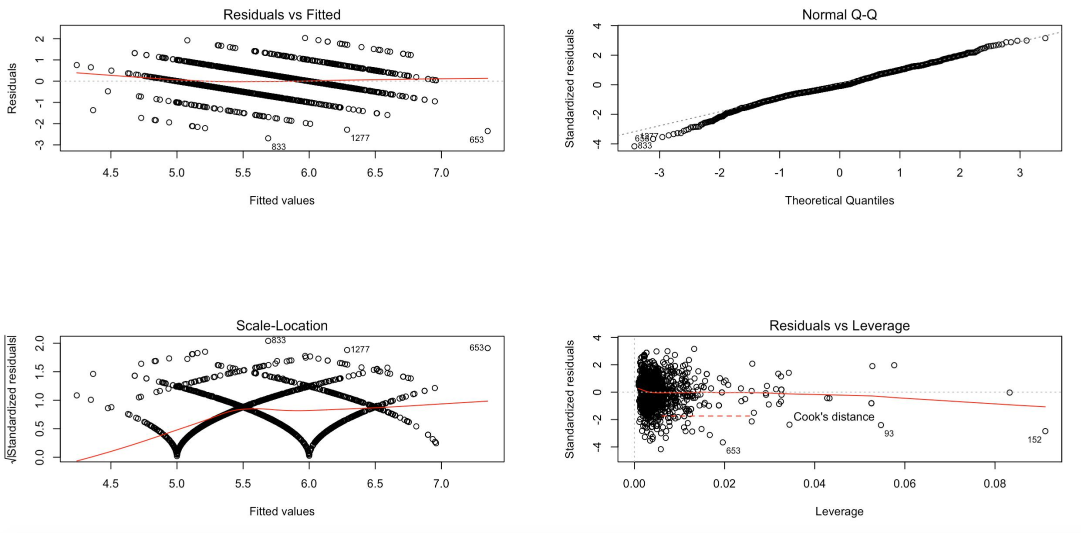
$$Y_{RED} = 4.441160 - 1.011272X_2 - 2.022234X_5 + 0.005148X_6 - 0.00335X_7 - 0.485986X_9 + 0.882581X_{10} + 0.289259X_{11}$$

$$Y_{WHITE} = 144.5 + 0.06934X_1 - 1.761X_2 + 0.07767X_4 + 0.02078X_6 + 0.002765X_7 - 145.1X_8 + 0.6689X_9 + 0.6201X_{10} + 0.197X_{11} - 0.0001X_6X_7$$

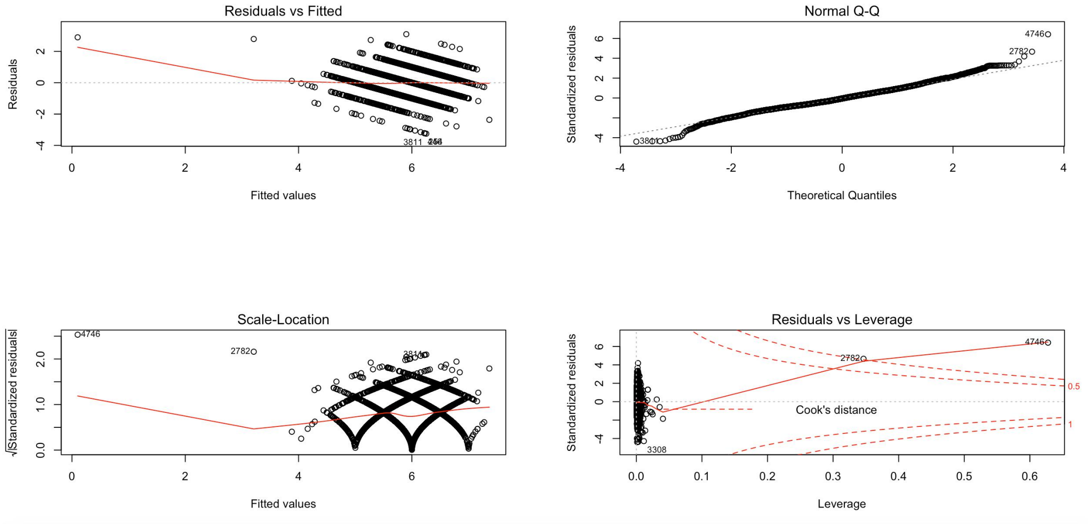
So far, we have established a set of relatively objective wine quality evaluation model.

We still need to test the statistical hypotheses in regression analysis. We use R studio to generate four charts which can assess the fitting of our model:

Red wine:



White wine:



As we can see from the above charts, the fitting effects of two polynomial regressions are close to ideal, which basically conform to the linear assumption, the residual normality and the homogeneity of variance.

Model checking

After completing the model, we need to know if the model is accurate, we extract a set of data from red wine and white wine respectively.

	Red wine	White wine
fixed.acidity	7.4	7
volatile.acidity	0.7	0.27
citric.acid	0	0.36
residual.sugar	1.9	20.7
chlorides	0.076	0.045
free.sulfur.dioxide	11	45
total.sulfur.dioxide	34	170
density	0.9978	1.001
pH	3.51	3
sulphates	0.56	0.45
alcohol	9.4	8.8
free.sulfur.dioxide: total.sulfur.dioxide	\	7650
The expert-graded wine quality	5	6
Predicted grade of wine quality	5.029776916	5.532074

(table 5)

Based on the above contents, we find that the errors are all within the acceptable range.

Conclusion

The final equations of assess the quality of red wine and white wine are:

$$Y_{RED} = 4.441160 - 1.011272X_2 - 2.022234X_5 + 0.005148X_6 - 0.00335X_7 - 0.485986X_9 + 0.882581X_{10} + 0.289259X_{11}$$

$$Y_{WHITE} = 144.5 + 0.06934X_1 - 1.761X_2 + 0.07767X_4 + 0.02078X_6 + 0.002765X_7 - 145.1X_8 + 0.6689X_9 + 0.6201X_{10} + 0.197X_{11} - 0.0001X_6X_7$$

As we can see from the two expressions above, there are some differences between the assess of red wine quality and white wine quality. After doing some research, we notice that there are two main reasons: the ingredients and the fermentation process, some can change by taking manual interventions but some can not. It Inspired us that we can take the expressions as a reference, although product labels won't be that detailed, we can offer some advice to consumers of buying wine:

1.Choose the wine with a relatively high pH;

2.Choose the wine with a relatively high alcohol strength;

3.For white wine, choose the one is relatively sweeter and clear.

appendix