Nara Shin

DS 312

Professor Jamie Eng

May 5, 2022

Finding the Best Equation for Deciding the Quality of Red Wine

This project is about deciding the most related X variables for red wine quality based on physicochemical tests to find the best equation that predicts Y variable. The data set used on the project was collected by UC Irvine Machine Learning Repository in 2009. This data set has total 1599 sets of observations. According to P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis who had conducted these tasks, the original datasets were created using red and white wine samples, but only red wine dataset was used in this project. Additionally, only physicochemical and sensory variables are available meaning that there is no data about grape types, wine brand or wine selling price to avoid of privacy and logistic issues that might be caused.

The attributes are originally eleven input attributes and one output attribute. Input variables based on physicochemical tests are these: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. The output variable is quality that is scale of 0 to 10. Based on these input attributes, one of them had to be converted to qualitative since at least one category variable is needed for this project. Out of all eleven attributes except output attribute, residual sugar was converted to category variable. As the average value of residual sugar was approximately 2.54 (rounded up from 2.53881), 0 is considered that residual sugar is less than or equal to 2.54, and 1 is considered that residual sugar is greater than 2.54. Based on these setting, multiple linear regression analysis had been done on Excel using Data Analysis Tool Pack.

obs	Y=quality	fixed acidity	volatile acidity	citric acid	chlorides	free sulfur dio	total sulfur di	density	pH	sulphates	alcohol	converted res	residual sugar
1	5	7.4	0.7	0	0.076	11	34	0.9978	3.51	0.56	9.4	0	1.9
2	5	7.8	0.88	0	0.098	25	67	0.9968	3.2	0.68	9.8	1	2.6
3	5	7.8	0.76	0.04	0.092	15	54	0.997	3.26	0.65	9.8	0	2.3
4	6	11.2	0.28	0.56	0.075	17	60	0.998	3.16	0.58	9.8	0	1.9
5	5	7.4	0.7	0	0.076	11	34	0.9978	3.51	0.56	9.4	0	1.9
6	5	7.4	0.66	0	0.075	13	40	0.9978	3.51	0.56	9.4	0	1.8
7	5	7.9	0.6	0.06	0.069	15	59	0.9964	3.3	0.46	9.4	0	1.6
8	7	7.3	0.65	0	0.065	15	21	0.9946	3.39	0.47	10	0	1.2
9	7	7.8	0.58	0.02	0.073	9	18	0.9968	3.36	0.57	9.5	0	2
10	5	7.5	0.5	0.36	0.071	17	102	0.9978	3.35	0.8	10.5	1	6.1
11	5	6.7	0.58	0.08	0.097	15	65	0.9959	3.28	0.54	9.2	0	1.8
12	5	7.5	0.5	0.36	0.071	17	102	0.9978	3.35	0.8	10.5	1	6.1
13	5	5.6	0.615	0	0.089	16	59	0.9943	3.58	0.52	9.9	0	1.6
14	5	7.8	0.61	0.29	0.114	9	29	0.9974	3.26	1.56	9.1	0	1.6
15	5	8.9	0.62	0.18	0.176	52	145	0.9986	3.16	0.88	9.2	1	3.8
16	5	8.9	0.62	0.19	0.17	51	148	0.9986	3.17	0.93	9.2	1	3.9
17	7	8.5	0.28	0.56	0.092	35	103	0.9969	3.3	0.75	10.5	0	1.8
18	5	8.1	0.56	0.28	0.368	16	56	0.9968	3.11	1.28	9.3	0	1.7
19	4	7.4	0.59	0.08	0.086	6	29	0.9974	3.38	0.5	9	1	4.4

Since this dataset has eleven variables, I decided to use the backward method because it can lessen the complexities of doing regression on Excel, which means that multiple regression had be conducted with all the eleven attributes instead of doing simple regression with each attribute individually. Before explaining the result, the term “sum of squared regression” means that the sum of the differences between the predicted value meaning quality in this project and the mean of the dependent variables meaning input attributes mentioned above. As the result of this multiple regression with 95% level of significance, sum of squared regression is roughly 375.771 out of 1042.165 (corrected total amount). To narrow down the variables that is not effective enough to predict the wine quality, “F test” is used. The numerator of “F*” is equal to sum of squared regression divided by degree of freedom (which means the number of coefficients). The denominator of “F*” is equal to Sum of corrected total, which is nearly 1042.165 minus Sum of squared regression, and then divide it by the numbers of observances minus one minus degree of freedom of sum of squared of regression; is 1587. Since the value of F* is greater than the value of F table, the hypothesis concludes HA that is at least one of β is not equal to zero.

ANOVA							
	df	*SS*	*MS*	*F*	*Significance F*	F table	1.79467068
Regression	11	375.771158	34.1610144	81.3535751	1.7564E-145	H0: all of β is equal 0	
Residual	1587	666.393945	0.41990797			HA: at least one of β is not equal 0	
Total	1598	1042.1651				since F* > Ftable, conclude HA	

To see which variable is not predictive enough, running 10 variables with every possible way is the next step. With the one that has the highest value of Sum of Squared Regression, calculate F* again. This time the numerator of F* is sum of squared regression given eleven variables must be subtracted by sum of squared regression given ten variables because what we like to know is how much density is telling from the numbers. Also, sum of corrected total needs to be subtracted by sum of squared regression given eleven variables for denominator because that tells you how much density is occupied purely. If value of F* is less than value of F table, then conclude with H0 which leads to drop the density.

ANOVA												
	df	*SS*	*MS*	*F*	*Significance F*	*ssreg(11)*	375.771158	0.21648191				
Regression	10	375.554676	37.5554676	89.4646711	2.335E-146		666.393945	0.41990797				
Residual	1588	666.610427	0.41977987				1587	0.51554609	F*	3.84732502	Ftable	
Total	1598	1042.1651						F>F*	conclude H0	drop density		
						H0: β_density = 0 given β other variables			HA: β_density is not equal to 0 give β other variables.			

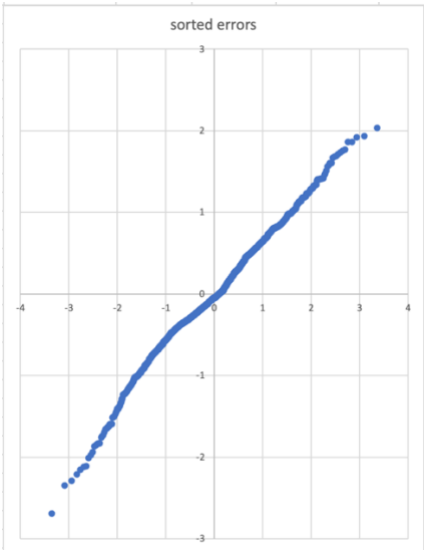
Repeating these processes until the value of F* is greater than value of F table, so that the equation can have the best X variables to predict Y variable. In this project, it ended up using 7 variables which are volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, pH, sulphates, and alcohol. Consequently, the best equation for deciding wine quality is $Y = \beta 0 + \beta vaXva + \beta chXch + \beta fsdXfsd + \beta tsdXtsd + \beta phXph + \beta sXs + \beta aXa + \varepsilon$.

Despite finding the best equation with the most effective X variables to predict Y variable, there is one more step to do in order to determine if this model adequately fits the data meaning if the equation follows these assumptions: homoscedasticity and normality. If these assumptions are violated, the model need to be improved. In regression model, the assumption of

homoscedasticity means that the variance of residuals is constant. Variance is a measurement of the spread between numbers and in a dataset (Hayes), and residual is a difference between measured value and predicted value. Thus, comparing the ratio between upper and lower standard residual values would tell if this model followed assumption of homoscedasticity. There is total 1599 sets of data, and the upper standard residual value is roughly 0.637 and the loser standard residual value is roughly 0.652. Dividing upper standard residual value by lower standard residual value equals to approximately 0.977 which is greater than 0.5 and less than 2. That is, the model follows assumption of homoscedasticity.

homoscadasticity	s1/s2	0.97709077
		0.5<0.977<2

Another assumption that should meet is normality of error distribution. The residuals of the model should be normally distributed. This also need to test if residual values are normally distributed. To check if the model is normally distributed, compare the inverse of the standard normal cumulative distribution with sorted errors of each observation. If the graph looks straight line, it means the model is normally distributed.



In conclusion, the model $Y = \beta 0 + \beta vaXva + \beta chXch + \beta fsdXfsd + \beta tsdXtsd + \beta phXph + \beta sXs + \beta aXa + \varepsilon$ is the best multiple linear regression models that meets two assumptions of regression analysis predicting red wine quality based on physicochemical and sensory variables.

Works Cited

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

Hayes, Adam. "Using the Variance Equation." *Investopedia*, Investopedia, 16 May 2022, https://www.investopedia.com/terms/v/variance.asp.