**Final report – Which white wine has the best quality?**

**103ROTC** - HAN DUCK SEO, JOO YOUNG CHOI, KANG SU LEE, YOUNG JE WOO, JIN UK LEE



**●Introduction/Background of Project and Research Question.**

Back then, Wine as regarded expensive and luxury item in korea but it becomes more available these days. Consumers nowadays seems to value the quality assurance tag of each wine. so it becomes important that how the quality of wines is evaluated. wine’s quality is evaluated by chemistry method and sensation method. pH, alcohol, and density is important element of chemistry method. wine is consist of water and alcohol mostly, and the others are sugar, organic acid. I’ll discuss which element makes wine’s quality better and by this analysis, I want to advise consumers which wine is good to drink and high quality.

Dataset is extracted from [Cortez et al., 2009] and these are composed of quality of wine and variable which influence on it. The quality variable is put red wine and white wine data. But we only analyze white wine data because each data has more than a thousand observations. There are too many variables to analyze. So we decided to analyze white wine data only.

In this data set, Researchers are putting the score of the quality of wine from 0 to 10. The rest of the independent variable's observations were judged to be made without stardardization because they are not beyond the excess range. The response variable is the quality of wine, and independent variable is wine's eleven physical-chemical elements. We use statistical methods to analyze data including regression, ridge, lasso, K-means clustering.

**●Statistical methods & results**

**1. Regression**

First, we divided dataset to training dataset and test dataset. And we set a model using traning dataset. Training dataset is 60% and test dataset is 40% of dataset. The result of regression and beta, p-value are like as following.

|  |  |  |
| --- | --- | --- |
| Intercept | 2.00E+02 | 6.44e-12 \*\*\* |
| fixed.acidity | 8.91E-02 | 0.001792 \*\* |
| volatile.acidity | -1.69E+00 | < 2e-16 \*\*\* |
| citric.acidity | 2.12E-01 | 0.084769 . |
| residual.sugar | 9.61E-02 | < 2e-16 \*\*\* |
| chlorides | -2.11E-01 | 0.769524 |
| free.sulfur.dioxide | 2.83E-03 | 0.007632 \*\* |
| total.sulfur.dioxide | 3.61E-04 | 0.467364 |
| density | -2.01E+02 | 9.56e-12 \*\*\* |
| pH | 8.86E-01 | 3.86e-10 \*\*\* |
| sulphates | 7.33E-01 | 1.55e-08 \*\*\* |
| alcohol | 1.33E-01 | 0.000278 \*\*\* |

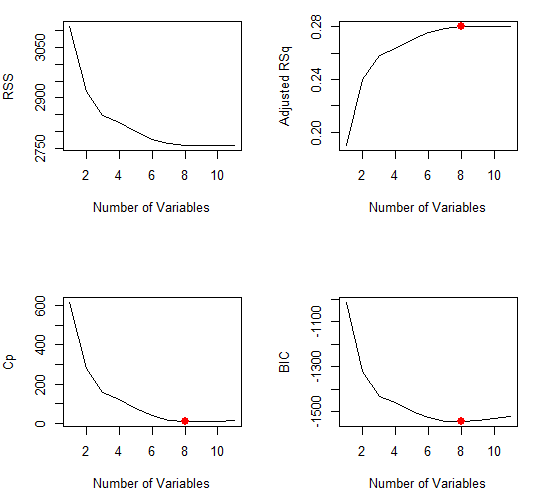
And Adjusted R-squared is 0.2805. It is appropriate considering it is a real data. F-test’s p-value is < 2.2e-16. It is very significant. The results of full model’s test MSE is 0.5834992. We could know about some variables (fixed.acidity, citric.acidity, residual.sugar, free.sulfur.dioxide, total.sulfur.dioxide, pH, sulphates, alcohol) are getting higher when the quality is getting higher and another variables (volatile.acidity, chlorides, density) are getting lower when the quality is getting higher. The Significant variables are most of them except 2 variables. But the interpretation of model like that. We are doubt the multicollinearity because there are similar variables. So we get a result by testing models using vif function.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| fixed.acidity | volatile.  acidity | citric.acid | residual.  sugar | chlorides | free.sulfur.  dioxide |
| 3.164199 | 1.122904 | 1.168842 | 15.673061 | 1.217758 | 1.804744 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| total.sulfur.  dioxide | density | pH | sulphates | alcohol |
| 2.302537 | 38.040474 | 2.336328 | 1.170145 | 10.670002 |

Normally, It is doubted about multicollinearity when the value is more than 10. 3 variables are doubted about multicollinearity. Therefore, we found better model by using best subset selection.

The next plot is about RSS, Adjusted RSq, Cp, BIC when the best subset selection’s result variable is 1,2,3,4,5,6.



We can see there are no difference when there are more than 8 variables. When there are 8 variables, Adjusted RSq is maximum. But Cp and BIC is minimum. So we decided the best model which has 8 variables. The Best model is like as following.

quality=1.541062e+02+6.810394e-02\*fixed.acidity-1.888140e+00\*volatile.acidity+8.284724e-02\*residual.sugar+3.349015e-03\*free.sulfur.dioxide-1.542913e+02\*density+6.942135e-01\*pH+6.285081e-01\*sulphates+1.931628e-01\*alcohol.

And vif of best model is likely as following.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| fixed.  acidity | volatile.  acidity | residual.  sugar | free.sulfur.  dioxide | density | pH | sulphates | alcohol |
| 3.018132 | 1.0435 | 9.401543 | 1.14383 | 34.294529 | 2.24232 | 1.166545 | 8.422527 |

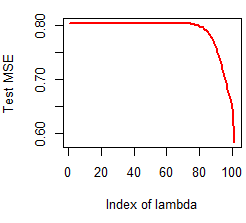
But the multicollinearity of density is still higher. So we eliminate density variable.

The final model is likely as following.

quality=1.541062e+02+6.810394e-02\*fixed.acidity-1.888140e+00\*volatile.acidity+8.284724e-02\*residual.sugar+3.349015e-03\*free.sulfur.dioxide+6.942135e-01\*pH+6.285081e-01\*sulphates+1.931628e-01\*alcohol.

**2. Ridge:**

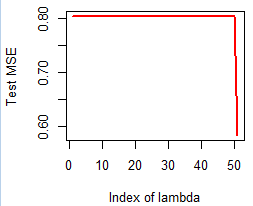
We used ridge to decreasing coefficient because there are too many variables. That is result below.

.

The hundredth value of lambda has the smallest test MSE. This the hundredth lambda is result of measurement according to r. the result is > l = which(mse == min(mse)) > ridge$lambda[l] [1] 0. In other words, lambda’s value is 0 and the best model is that does not reduce some coefficient. Final model through ridge likes that full model. Therefore, there is no need to use this ridge.

**3. Lasso:**

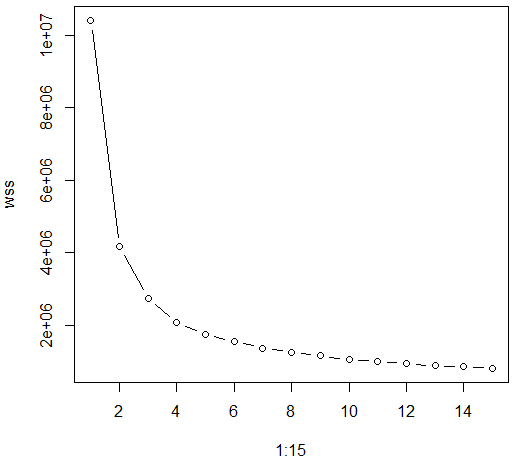
We used also lasso to decreasing coefficient. Lasso is similar ridge and prediction but interpretability of lasso is better. Because our purpose of analysis is prediction, there is no necessity for using lasso. But we analyzed lasso for assuring that follow similar result. The result is like that below.



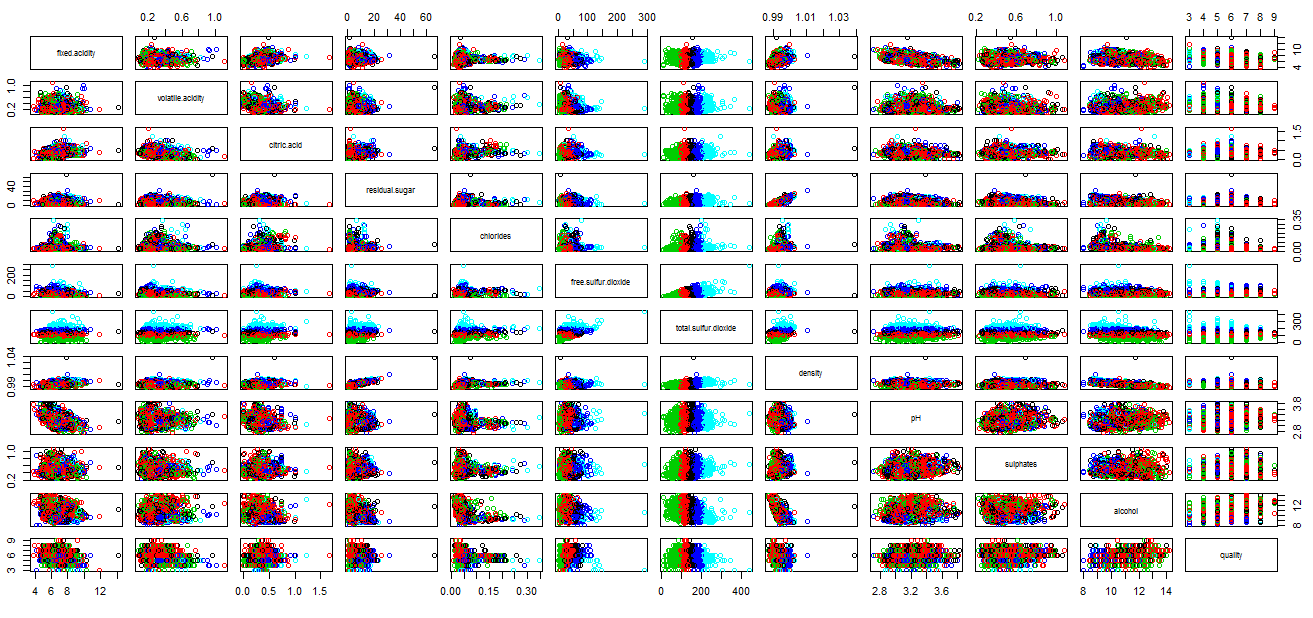
This time, we reduced section of lambda by 50. When lambda is 50, that is to say the last, test MSE has minimum value. Following this result using r is 0 like ridge. In other words, the best model is that does not reduce coefficient likes ridges. Final model is same full model. Therefore, there is also no need to use lasso.

**4. K-means clustering:**

Lastly, we progressed clustering. It's the method how to grouping data when we don't know the right answer on data or don't have one on unsupervised learning method. First, we drew plot based on the wss for deciding group’s number. The plot is that below.



Examining plot, there are no gap after 5 groups. So, we set up 5 group and progress analysis. Groups to be divided describe plot

.

Because data’s number is 4898, we are hard pressed to figure out plot. we organized clustering result by graphs.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | fixed.acidity | volatile.acidity | citric.acid | residual.sugar | chlorides |
| Group1 | 6.960524 | 0.2855388 | 0.3537160 | 8.824018 | 0.05114804 |
| Group2 | 6.840321 | 0.2724726 | 0.3359326 | 7.006975 | 0.04732367 |
| Group3 | 6.813370 | 0.2799025 | 0.3158357 | 3.450557 | 0.04015042 |
| Group4 | 7.010969 | 0.3073980 | 0.3557908 | 10.033801 | 0.05228571 |
| Grop5 | 6.777090 | 0.2700066 | 0.3230678 | 4.734200 | 0.04193153 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | free.sulfur.dioxide | total.sulfur.dioxide | density | pH | sulphates |
| Group1 | 47.02971 | 179.05690 | 0.9959135 | 3.183112 | 0.5075831 |
| Group2 | 37.70415 | 145.32367 | 0.9944275 | 3.198770 | 0.4855643 |
| Group3 | 18.86212 | 77.53482 | 0.9918326 | 3.175864 | 0.4691086 |
| Group4 | 55.29847 | 221.74617 | 0.9968072 | 3.178265 | 0.5180357 |
| Group5 | 28.24753 | 113.13989 | 0.9927783 | 3.191257 | 0.4843779 |

|  |  |  |
| --- | --- | --- |
|  | alcohol | Quality |
| Group1 | 9.831101 | 5.629406 |
| Group2 | 10.397542 | 5.938088 |
| Group3 | 11.255687 | 5.903900 |
| Group4 | 9.541582 | 5.522959 |
| Group5 | 10.959480 | 6.069124 |

Comparing Group5 that quality is the highest and Group4 that quality is the lowest, this quality is better that is higher alchol, pH and lower suphates, density, total.sulfur.dioxide, free.sulfur.dioxide,chlorides, residual.sugar, citric.acid, volatile.acidity and fixed.acidity. But when compared other groups, these things always are not content Therefore, we know this clustering is failure. Also we know that it’s different from the regression result. It wasn’t easy to analyze by Clustering because of the number of data is too large to do so.

**●Discussion**

1. In general, evaluating the taste of wine is widly depended on expert evaluation better than statistic technics. Commonly referred to as ‘Sommelier’, they are expert in evaluation wine. Therefore, comparing Sommelier’s assessment and statisic analysis, we can derive more meanigful results. However, when two different part’s results are direct-opposed, which ones are more credible?

2. Our one of the biggest problem that we may encouter when we perform this project is that can we quantify the different than ours. especialy, wine makes special air between couple on anniverary day. it means that the taste of wine differentiate as who i drink with and where i drink. Can we deny those parts?

**● Conclusions and Future Directions**

In doing regression, it is appropriate considering it is a real data. Double checking the regression page, F-test’s p-value is < 2.2e-16. It is very significant. The results of full model’s test MSE is 0.5834992. We could know about some variables (fixed.acidity, citric.acidity, residual.sugar, free.sulfur.dioxide, total.sulfur.dioxide, pH, sulphates, alcohol) are getting higher when the quality is getting higher and another variables (volatile.acidity, chlorides, density) are getting lower when the quality is getting higher. Therefore, if we check these variables, we can find delicious wine.

Now, both consumers and producers can do acts how they should behave.

wine producers produce a wine which is following statistic results, and consumer buy the wine which is also following statistic results. It can be used as an important indicator in the wine market.

**● References**

**The source of data set** is Paulo Cortez, University of Minho, Guimarães, Portugal, <http://www3.dsi.uminho.pt/pcortez/Home.html>, A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal @2009

**Data Set URL** : <http://archive.ics.uci.edu/ml/datasets/Wine+Quality>