MSCI 718-Assignment 3- Wine Data

Shabnam Hajian & Dina Haji Zeynaly Biooky

2020-03-31

Summary of the dataset

This data has been collected from red and white Vinho Verde wine samples from the north of Portugal to quality assessment. Wine qualification is based on the physicochemical tests from laboratory tests plus sensory tests which is more based on human tastes with the complexity of it (as human concepts always have). Thus, the relationships between the physicochemical and sensory tests are complex and still not fully understood.[1] The data comes from two separate datasets that included related data of two different wine colours, red and white. The *red* dataset contains *1599* observations and the *white* one has *4898* observed data. Also, both data sets have *12* variables before any changes.

One of the most important variables is *quality*. This variable is the outcome variable and its data is a score between 0(very bad) and 10(very excellent) resulted as a median of at least 3 evaluations made by wine experts. In our dataset, quality column has a **minimum=3**, **maximum=9**, **median=6**, and **mean=5.818**. Data is integer and approximately normal. Other physicochemical variables are as below: "*alcohol"* which is between 8 and 14.9 and skewed to left. "*sulphates"* which acts as an antimicrobial and antioxidant and should be in the limited amount. "*volatile acidity"* is the amount of acetic acid in wine, it could cause an unpleasant taste in a higher amount. "*citric" acid* that found in small quantities to add freshness and flavour to wines. "*residual sugar"* is the amount of sugar remaining after fermentation stops and its possible range is between 1 and 45 grams/litre. Our dataset's "*residual sugar*" mean is 5.44, median is 3 means most of numbers are small. "*fixed acidity"* is the combination of 4 to 5 acids. Our "*fixed acidity*" data is skewed to left with the range between 3.8 and 15.9. "*Chlorides"* shows the amount of salt in the wine and the majority of our data is below 0.05. The "*free sulfur dioxide"* is a form of SO2 that exists in wine to prevents microbial growth and the oxidation of the wine and usually kept below 150. "*total sulfur dioxide"* is the amount of free and bound forms of S02 and this data looks bimodal. "*density"* and "*pH"* are our last variables which look approximately normal in this dataset with the mean of 0.99 and 3.21 respectively.

In this report, we try to model a multiple linear regression to predict the quality of the wine based on the most related and available variables from the datasets. To do this first we need to make some assumptions. For the purpose of simplicity and to be able to apply multiple linear regression in our analysis, we assume that the relationship between the quality as an outcome and its predictor variables are linear and the output data is quantitative, continuous, and unbounded (although we know it is not). Second, though the red and white wines are quite different and might have different behaviour, we do not perform separate analyses for them and assume approximately the same qualifications scale for both wine colours. Third, we use the mean for quality prediction; however, the median might be a more realistic approach.

Analysis

In this analysis, we are trying to predict wine's quality based on its influential factors by using the multiple linear regression model. To do so, we will do the following steps.

Tidy Data

After getting a general idea about our data set and variables, we understand that we need to make some changes into our data to make it "Tidy". First, it seems that each observation is in one row, each variable has one separate column and each value has its own cell. However, data are in two separate files, we need to combine these two files into one to make our analysis easier. Before combining two files, we need to make sure that we will not miss information about the color of our final product. Therefore, we add a new variable and call it "color" and fill it "red" or "white" based on the information we have from the dataset. After these transformations, we need to look for missing data. Since there is no missing data, we will check

[1] 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.

for outliers. In this step, we will use visual scatter plots and check the minimum and maximum of the data points since our main purpose here is to check data for any data entry errors. The main outliers and influencers will be specified after we build our model. Based on the usual and possible ranges for each of these parameters, we would see some areas of concern. First, we should pay attention to *residual sugar*. There is a data over 45 gr which could be rare, and it has a considerable distance from the rest of our data. The second variable is *free_sulfur_dioxide* which again has data far from the rest of the sample. The last variable is *density* as there are two data relatively far from the rest of our data. Thus, we consider these points as strange observations and because we do not have any access to the original data to check these observations, we would eliminate theses cases from our data set for now. This means that we have a new sample with 6,493 row which has 4 observations less than the original file. Now we can conclude that our data is tidy and clean with no data entry error and ready to model and analyze. As we said, we will control outliers and influencers again after we build our model.

Selecting predictor variables

Based on the P. Cortez et al., we know that the most important factors in the quality of the wine are *sulphates*, *alcohol*, and *volatile acidity*. Also, we know that *citric acid*, and *residual sugar* play an important role in predicting the quality of wines although the amount of this impact is different in two different wine colors. Moreover, this study emphasizes the wine color is also important. As we mentioned, we will not consider two separate models, but we will use color as a categorical variable in our model. Thus, we use the **Hierarchical** method and based on these six predictors, model a multiple linear regression with a categorical variable. From the model results, we conclude that *sulphates*, *alcohol*, *volatile_acidity*, *citric_acid*, *residual_sugar* and, also *color* as a categorical variable explain approximately 30% of the variance in predicting wine quality(**R-Square=0.284**) and **adjusted-R square=0.2834**. In addition, we can see the influence of these variables on future wines' quality is significant for all 6 predictor variables and the intercept with 95% confidence based on the p-values.

As the next step, we want to use **Forced-entry** method to check this hypothesis: "Is any of the other predictor variables have a significate impact on predicting outcome". After running this model, results illustrate new **R-square=0.2995** and the new **adjusted R-square=0.2983**. We expected that based on increasing the number of predictor variables, this number decrease, although it shows increase. To check difference between to model, we use ANOVA test. This test is based on the **null hypothesis** as: *There is no difference between these two models in predicting outcome.* and the **alternative hypothesis** as: *These two models are different in predicting outcome variable.* Based on the p-value of the ANOVA test(p-value= 2.2e-16), we could say that this model is significantly different from the first model to predict wine quality with 5% level of significate. Also, as the R- square and adjusted R-square are improved, we could conclude that this model is better than the first one and we will continue with this model.

```
##                                               2.5  %            97.5  %
##    (Intercept)                             9.436276e+01      1.557545e+02
##    sulphates                               5.978267e-01      8.974904e-01
##    alcohol                                 1.596661e-01      2.374045e-01
##    volatile_acidity                       -1.646588e+00     -1.328322e+00
##    citric_acid                            -2.228795e-01      8.900149e-02
##    residual_sugar                          5.554561e-02      8.013936e-02
##    fixed_acidity                           7.034437e-02      1.357044e-01
##    chlorides                              -1.388976e+00     -8.071237e-02
##    free_sulfur_dioxide                     4.176786e-03      7.227934e-03
##    total_sulfur_dioxide                   -1.982012e-03     -7.101483e-04
##    density                                -1.560586e+02     -9.370735e+01
##    pH                                      3.876276e-01      7.525833e-01
##    colorIf_red                             2.994260e-01      5.310129e-01
```

From the output of coefficients of this model, we can see except *citric_acid,* all other variables coefficient and intercept are significate in in model. The *citric_acid* has a relatively big p-value(0.4001) also, as we can see above, the coefficient interval contains zero; this means that this variable does not show a significate impact on outcome. The important note about *citric_acid* is we knew based on the study of P. Cortez et al., this variable is an important predictor in the quality model, and we can not eliminate this variable. This contradiction might exist as a result of between variables' effects. To find a suitable combination of predictor variables, we have to use the **all-subset** method to test all 32 possible combinations and find the best one for our model, but in this assignment, we accept this model with **R-square=0.2995** and **adjusted R-square=0.2983** and continue.

Outliers and influential cases

As we know *Outliers* and *influential* points can have an impact on your model. Now, as we have a specific multiple linear regression model, we could have more justification to remove some data to create a more accurate model. Based on the concept of outliers, we need to look for some residuals that are far from the 0. Here we used *standardize residuals* since they are unitless and easier to compare and 95% confidence. Results show 5.6% (361 data out of 6,493) of observations are out of two standard deviations far from the mean. Also, there are 53 observations out of 3 standard deviation that might create some problem.

Since *Outliers* may not influence our model, we need to use another measure to find *influential* observations as *Cook's distance.* We know that Cook's distances greater than 1 are a cause for concern. However, the maximum of Cook's distances is **0.023** which is considerably less than 1 and we can conclude that there is no influential observation in these datasets for this model. These numbers mean our model robust and everything is good.

Checking assumptions of multiple linear regression

Now as we built the multiple linear regression model, we need to check assumptions.

1. Quantitative or categorical predictor variables and quantitative, continuous, and unbounded outcome: Although our outcome variable could not meet this essential characteristic, based on what we said at the beginning, we assume this assumption is **met**.

2. non-zero variance: we could easily check this assumption by calculating the variance of each variable.

```
##             sulphates   alcohol   volatile_acidity      residual_sugar   fixed_acidity
## Variances: 0.02214408   1.422331     0.02706062            21.87744         1.681387
##            free_sulfur_dioxide   total_sulfur_dioxide       density            pH
## Variances:     304.8586              3179.067            8.620006e-06       0.02585492
```

Honestly, there are some concerns, but let's move on.

3. no perfect multicollinearity: we use tolerance = (1/VIF) to check multicollinearity.

```
##  sulphates   alcohol    volatile_acidity   citric_acid   residual_sugar  fixed_acidity
## 0.63656095  0.14726280    0.46179469       0.61625103     0.09565773      0.17622722
##  chlorides  free_sulfur_dioxide total_sulfur_dioxide  density      pH        color
## 0.60222834  0.44600642           0.24614167          0.03777178  0.36757194 0.12713039
```

We can see some numbers are even below 0.1 which might create some problem. Also, we also check to mean of *vif* that is 6.1087282 and below 10.

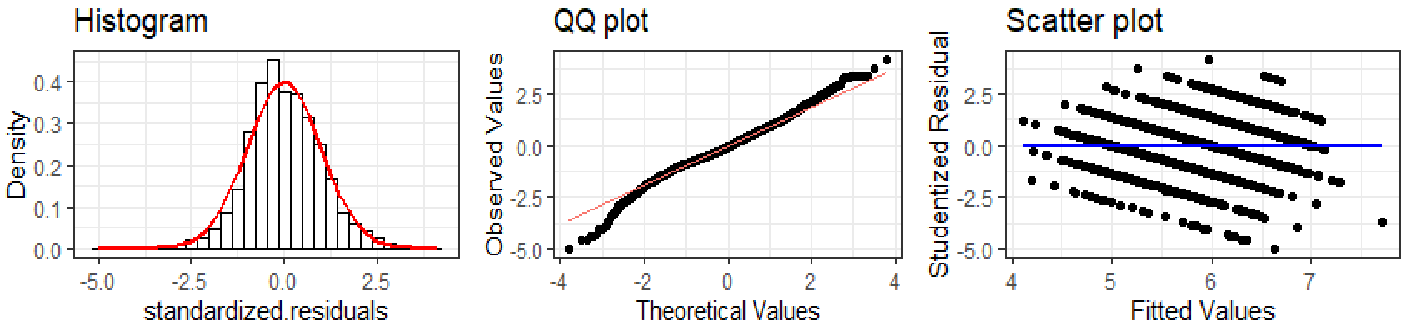
4. predictors are uncorrelated with external variables: Actually, we could not make sure about this, but we assume this assumption is **met**.

5. residuals are homoscedastic, independent, and Normal: To test independency of the residuals, we use Durbin Watson Test:

```
## lag  Autocorrelation  D-W Statistic  p-value
## 1       0.1772937       1.645394        0
##  Alternative hypothesis: rho != 0
```

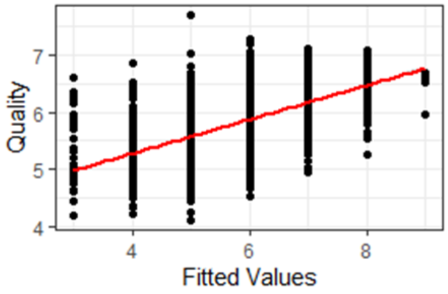
The test results indicate d is very close to 2(as the threshold). Thus, we do not reject the null hypothesis that the errors are significantly independent and continue with the assumption of independence residuals at the 5% level of significance.

For homoscedasticity and linearity, since our sample is big, we preferred to use visual plots instead of the statistical tests.



From the "Histogram" plot, we can observe that the residuals are not normal, and its distribution is skewed to the left. Also, if we look at the "QQ plot" we can see that this visual guess from the histogram was correct as points do not lie along the red line. The last plot is "Scatter plot". This plot shows that the residuals are approximately homoscedastic. We use color for different wines' color to show there is no difference between these two wine types.

Based on all the above, we can conclude that assumptions of *Indipendency* and *Homoscedasity* are **met** but the residuals are not *Normal* and this assumption **violated**.



6. linearity of outcome variable: Lets first use a plot!

Frankly, we know that this is not a good shape as we expected. The point is "quality" as the outcome variable is not as it has to(quantitative, continuous, and unbounded), but based on first assumptions, we used linear regression(although we knew it was not a good choice) to model the output! As a result, let's accept this assumption is **met**.

Now after checking model assumptions, we can see that some of these assumptions are violated and the model could not robust it. Thus, we conclude that even if we accept this model for this sample, we could not generalize it.

Conclusion and potential follow up

In this report, we tried to use multiple linear regression to predict wine quality based on the 11 predictor variables from our sample data and the wine color based on some assumptions we made in the first part which explained nearly 30% of the variance. Thus, the model seems not an accurate prediction of wine quality. This could be as a result of our first assumption about the outcome variable. Obviously, other models with different assumptions might be a better fit for this data. Another possible improvement could happen if we apply all subset approach and check all possible combinations to select predictor variables. Also, from the raw data, we realized that white and red wines are considerably different. As a result, it would be a better idea to generate two separate models for these two types of wine instead of considering it only as a categorical variable in our model.