

Intelligent Data Analysis - Homework 1.2

Panagiotis Michalopoulos, Javier de la Rua, Michail Gongolidis,
Ignacio Rodriguez, Daniel Minguez

October 24th, 2018

1 Basic exploration of a dataset

1.1 Introduction

We have chosen the wines dataset, this dataset has 11 variables related to tests on different wines plus a categorical variable grading the wine quality by experts and a binary variable indicating the type (white or red):

- **Fixed acidity:** related to fixed acids or metabolic acids. They are produced, for example, from an incomplete metabolism of carbohydrates, fats, and proteins.
- **Volatile acidity:** form of wine spoilage. Unit of measurement not appreciated.
- **Citric acid:** is a weak organic acid. Unit of measurement not appreciated.
- **Residual sugar:** percentage of sugar not fermented.
- **Chloride:** Cl-, high concentrations makes the wine salty (pH >7).
- **Free sulphur dioxide:** sulphur dioxide is used as a stabilizer in wine production to prevent undesired biochemical processes in the must and the final product.
- **Total sulphur dioxide.**
- **Density:** mass/volume.
- **pH:** (acid, salinity meter).
- **Sulphates:** turns into sulphur dioxide (S-).
- **Alcohol percentage.**
- **Quality:** rated from 0 to 10.
- **Type:** (1= white wine or 2).

1.2 Descriptive statistics

1.2.1 Questions

- a) Choose a quantitative variable and explore its distribution in terms of descriptive measures of center, dispersion, skewness and kurtosis. Is a normal model a plausible one for its distribution? If the answer is no, can you think of a transformation of the variable that improves normality. Are there any outliers?

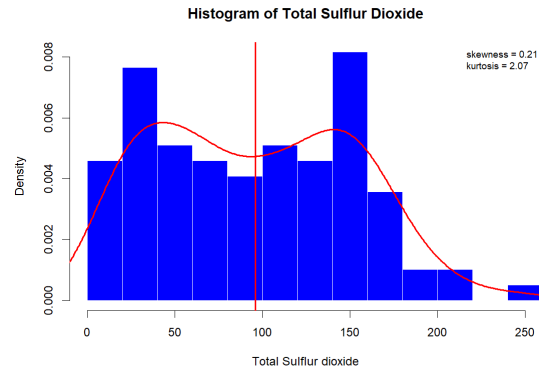
We have choose to explore the quantitative variable “Total Sulfur Dioxide”, and we have calculated the statistics related with center, dispersion, skewness and kurtosis.

Summary						
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	
11.00	46.25	98.00	96.01	142.75	245.00	
Sample variance						
[1] 3166.34						
Standard deviation						
[1] 56.27024						
Coefficient variation						
[1] 58.60861						
Quantile						
0%	25%	50%	75%	100%		
11.00	46.25	98.00	142.75	245.00		

The skewness is 0.2167518. This value implies that the distribution of the data is slightly skewed to the right or positive skewed.

It is skewed to the right because the computed value is positive, and is slightly, because the value is close to zero. For the kurtosis, we have 2.073274 implying that the distribution of the data is platykurtic, since the computed value is less than 3.

By checking the histogram of the variable, its easy to check that it seems to not follow a normal distribution because it shows two different peaks (bimodal):

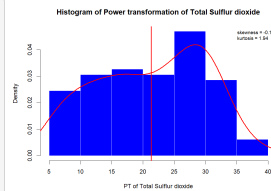


We also have tested normality (check shiny App) with the Anscombe-Glynn Kurtosis-based test and D'Agostino skewness-based test. The kurtosis test rejects normality in part because of the bimodality that we can see in the distribution, on the other hand, the skewness is closer to a normal distribution. Finally the jarque bera test has a p-value greater than 0.05 but still small (0.11). In order to improve the normality of the data, we have chosen to apply the the Box-Cox transformation:

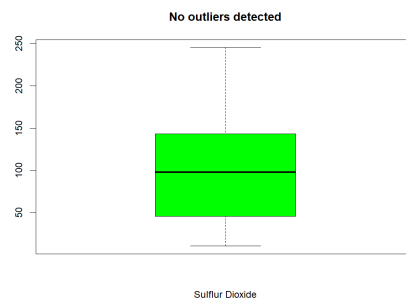
```
bcPower Transformation to Normality
      Est Power Rounded Pwr Wald Lwr Bnd Wald Upwr Bnd
wines$TSO2  0.5788      0.5  0.2917      0.8658

Likelihood ratio test that transformation parameter is equal to 0
(log transformation)
      LRT df      pval
LR test, lambda = (0) 17.095 1 3.5556e-05

Likelihood ratio test that no transformation is needed
      LRT df      pval
LR test, lambda = (1) 7.695133 1 0.005537
```

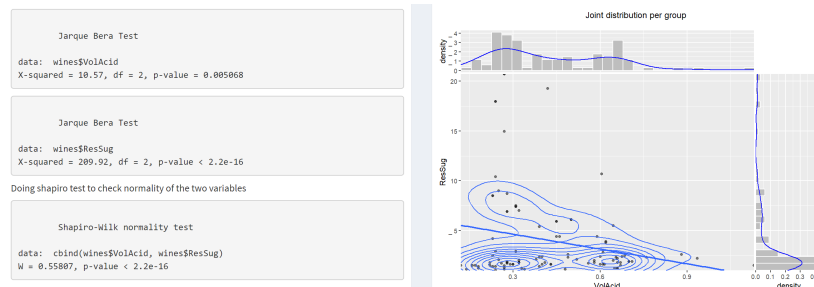


After applying the transformation and reapplying the test, we can see that there is not significance improvement regarding normality (check shiny App for the exact parameters of the tests). It could be because this variable is explained by others variables in the dataset. Finally, by using a boxplot graph, we have checked that there are not outliers in the data:

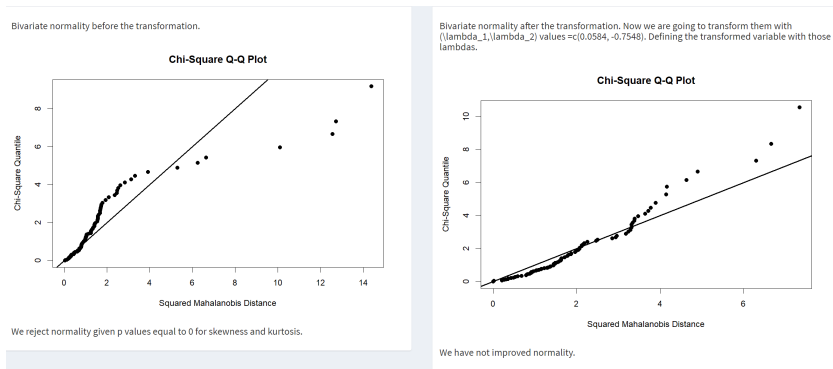


b) Choose two quantitative variables and describe its joint bivariate distribution. Does it seem to be Normal? Are there any outliers?

For this part, we have chosen the variables “Volatile acidity” and “Residual sugar”. As a first step we have tested if any of the variables by their own follows a normal distribution, after that we have applied the Shapiro Wilk test for bivariate data. We have seen that the data does not follow a normal distribution:



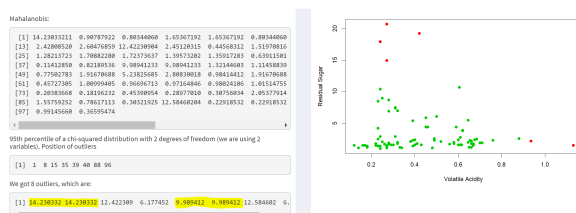
Then we have applied the box-cox transformation to the bivariate data (check shiny app), we can see that neither the transformation follows a normal distribution by following the fact that the mahalanobis distance of a bivariate normal distribution follows a Chi-Square distribution (mardia test):



In the qqplot for the original data we can appreciate that the distribution of the distance seems to be bimodal and right-light tailed compared with the chi-squared distribution. On the other hand, the transformed data seems to have heavier right tails.

Finally, we check for outliers by using also the mahalanobis distance and taking all and comparing the values that take the chi squared distribution with 2 degrees of freedom and 95% quantile.

We can see that it seems to be 8 outliers, with two pairs of two observations with the same Mahalanobis distance (and the same values), that is why in the scatterplot we can only see six red dots:



c) Choose a subset of 4 or 5 quantitative variables and explore linear relationships

We have chosen 5 quantitative variables which are *Residual sugar*, *Chlorides*, *Sulphites*, *pH* and *Alcohol percentage*.

2 Permutation tests

2.1 Choose a subset of 4 or 5 quantitative variables and explore linear relationships

a) R matrix of pairwise correlations

The following is the matrix of partial correlations:

Is it possible to appreciate:

- Moderate direct relation between Chlorides and Sulfites which makes sense as are used together for taste stabilizers.
- Negligible inverse relation between Residual sugar and pH levels, which tells us that the more sugar molecules rest in the wine, the pH will descent making the wine more acid. (The same happens in our mouth when we eat candies)
- Negligible direct relation between Alcohol percentage and pH, which can be explain with the logic of the point above, in the other words, more alcohol, less sugar so pH increases (less acidity).
- Negligible inver relation between chlorides and alcohol, as acyl chlorides consumes alcohols to produce esters.

b) Matrix of partial correlations

Partial correlation between 2 given variables, removes the external influences of the rest of the variables, by comparing this matrix with the R matrix we can reveal hidden relations between 2 variables alone, e.g. in our example we can see that the relations commented before are only enhanced so there are not hidden relations revealed. The Pearson coefficient between chlorides and alcohol percentage is still on the same value. The relationship between Chlorides and Sulfites has decreased, but Alcohol and Residual Sugar is enhanced which makes the most sense as the fermentation of sugar molecules creates ethanol molecules that increase the Alcohol percentage.

c) Coefficient of determination (function `r2multv()` we define in R)

Chlorides (chlor) is the best linearly explained by the others ($R^2 = 0.372$), followed by Sulphates (S, $R^2 = 0.308$). The worst linearly explained by the others is pH (pH, $R^2 = 0.129$).

d) The determinant of R (correlation matrix) as an overall measure of linear relationships.

The determinant of R is:

A non-zero $|R|$ indicates that there are not strong correlations among the variables and so there are no possible linear combinations.

e) An eigenanalysis of matrix R, looking for really small eigenvalues.

Looking at the eigenvalues we observe that Sulfites, pH and Alcohol percentage have eigenvalues closer to zero.

Lets analyze their eigenvectors:

- [3] For Sulfites, assuming all values are close to 0 except the third -0,968 we can conclude that this variable has very small variance, it is almost constant through the observed data.
- [4] For pH, assuming all values are close to 0 except the forth -0,978 we can conclude that this variable has very small variance, it is almost constant through the observed data.
- [5] For Alcohol %, assuming all values are close to 0 except the second -0,989 we can conclude that this variable has very small variance, it is almost constant through the observed data.

- 2.2** Repeat the analysis deleting the values for three customers that left a tip greater than 30% of the bill. These generous customers seem to be outliers.