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 if it is white wine or 2 if red wine).

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 chosen 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
```

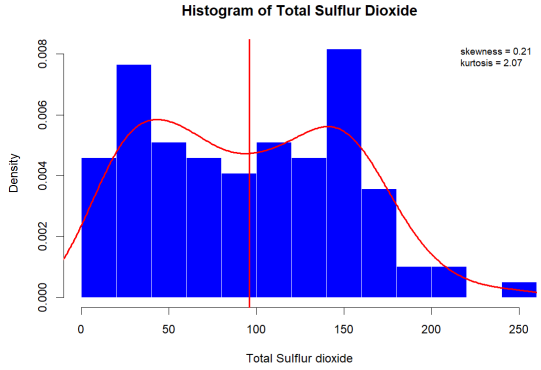
Quantile

```
    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):

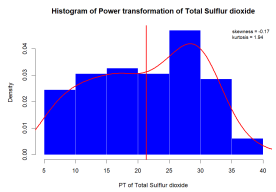


We also have tested normality (check shiny App) with the Anscombe-Glynn Kurtosis-based test and D'Agostino skewness-based test. The kurtosis test rejects normality in part because of the bimodality that we can see in the distribution, on the other hand, the skewness is closer to a normal distribution. 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 Upr 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:

```
        Jarque Bera Test

data:  wines$VolAcid
X-squared = 10.57, df = 2, p-value = 0.005068
```

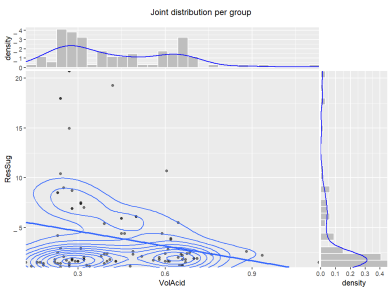
```
        Jarque Bera Test

data:  wines$ResSug
X-squared = 209.92, df = 2, p-value < 2.2e-16
```

Doing shapiro test to check normality of the two variables

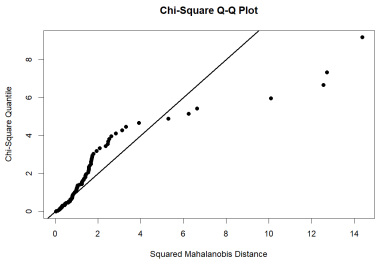
```
        Shapiro-Wilk normality test

data:  cbind(wines$VolAcid, wines$ResSug)
W = 0.55807, p-value < 2.2e-16
```



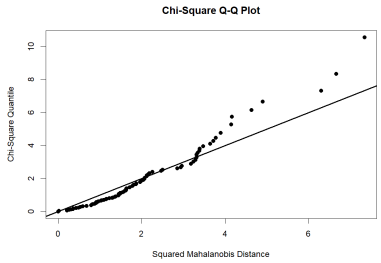
Then we have applied the box-cox transformation to the bivariate data (check shiny app), we can see that neither the transformation follows a normal distribution the fact that the mahalanobis distance of a bivariate normal distribution follows a Chi-Square distribution (mardia test):

Bivariate normality before the transformation.



We reject normality given p values equal to 0 for skewness and kurtosis.

Bivariate normality after the transformation. Now we are going to transform them with (\lambda_1,\lambda_2) values =c(0.0584, -0.7548). Defining the transformed variable with those lambdas.

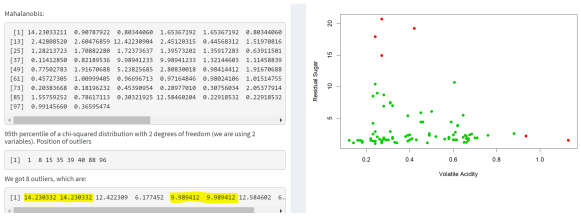


We have not improved normality.

In the qqplot for the original data we have appreciated 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 have checked for outliers by using also the mahalanobis distance and taking all and comparing the values that takes the chi squared distribution with 2 degrees of freedom and 95% quantile.

We have seen 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:

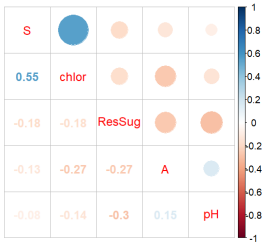


c) Choose a subset of 4 or 5 quantitative variables and explore linear relationships

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

- **R matrix of pairwise correlations**

The following is the matrix of partial correlations (check the shiny app for more examples):

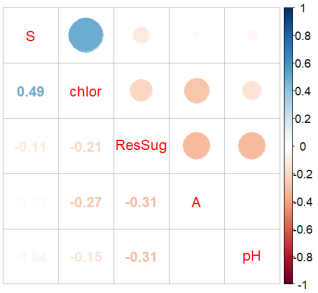


It is possible to appreciate a moderate direct relation between Chlorides and Sulfites and a neglible inverse relation between Residual sugar and pH levels, which tells us that the more sugar molecules rest in the wine, the pH will descent making the wine more acid. The rest of the relations are weaker, there is a neglible direct relation between Alcohol percentage and pH, more alcohol, less sugar so pH increases (less acidity) and finally a Neglible inver relation between chlorides and alcohol, as acyl chlorides consumes alcohols to produce esters.

- **Matrix of partial correlations**

Partial conrrelation between 2 given variables, removes the external influences of the rest of the variables. The coefficients seems to be still on the same value for

all the variables. The relationship between Chlorides and Sulfites has decreased, but Alcohol and Residual Sugar is enhanced which makes the most sense as the the fermentation of sugar molecules creates ethanol molecules that increase the Alcohol percentage.



- **Coefficient of determination (function r2multv() we define in R)**

Chlorides (chlor) is the best linearly explained by the others (R2 = 0.372), followed by Sulphates (S, R2 = 0.308). The worst linearly explained by the others is pH (pH, R2 = 0.129).

```
   ResSug     chlor         S        pH         A
0.2330976 0.3725142 0.3086383 0.1298088 0.1749346
```

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

The determinant of R is 0.4817213 (check shiny app). A non-zero $|R|$ indicates that there are not strong correlations among the variables and so there are no possible linear combinations.

- **An eigenanalysis of matrix R, looking for really small eigenvalues.**

We have calculated the eigenvalues and eigenvectors of the matrix with columns $[ResSug, chlor, S, pH, A]$. Looking at the eigenvalues we have observed that Sulfites, pH and chlor percentage have eigenvalues closer to cero.

```
eigen() decomposition
$values
[1] 19.659770920  0.823882540  0.039743582  0.019419378  0.002018158

$vectors
             [,1]        [,2]        [,3]        [,4]        [,5]
[1,]  0.998132330 -0.05926862 -0.009337335 -0.011222871 -0.002438107
[2,] -0.002259345  0.02114259 -0.144683299  0.022632753 -0.988990590
[3,] -0.008314147  0.04517882 -0.968465547 -0.202262448  0.138036719
[4,] -0.010024221 -0.01230237  0.196646221 -0.978997041 -0.051412297
[5,] -0.059641741 -0.99691913 -0.048829350  0.004062197 -0.013939432
```

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 chlor %, 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 Permutation tests

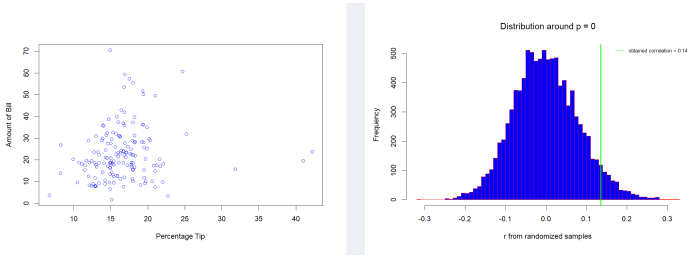
With Permutation testing the reference distribution is generated from the data themselves, instead of comparing the actual value of a test statistic to a standard statistical distribution. Permutation provides an efficient approach to test when the data do not conform to the distributional assumptions of the statistical method one wants to use (e.g. normality).

The null hypothesis is that correlation is equal to 0. This means that there is no linear relationship between the two variables. If that is true, then any of the Y observations is just as likely to appear with any of the X's. In other words, Yi is just as likely to appear with Xi as it is to appear with Xj, i$\neq$j.

2.1 Choose variables Bill (amount of Bill) and PctTip (tip amount as percentage of the bill) to analyse their linear dependency through Pearson's correlation coefficient. Just looking at the scatterplot, it is hard to tell whether this coefficient is significantly different from zero. Conduct a permutation test to test the null hypothesis that the correlation coefficient is 0 vs the alternative that it is different from 0. Run R = 10000 simulations.

Some economists have theorized that people tend to reduce the tip percentage when the bill gets large, but it could also be the other way around, customers might be more generous when eating in larger groups, thus spending more money, due to peer pressure. We have used the RestaurantTip data to see if there is evidence to support either theory, or perhaps there is no consistent relationship between the size of the bill and percent tip.

The scatter plot doesn't help us to make clear if the correlation coefficient is significantly different to zero.

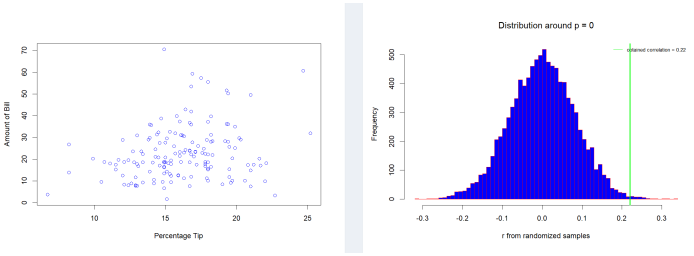


To figure out we have conducted a permutation test between the two corresponding variables. In the following histogram we can visualize the distribution around 0 of the random correlation coefficient computed during the 10000 simulations.

The permutation test results to a p-value that fluctuates around the significance level of 0.05, which leads us to some evidence against the null hypothesis.

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.

We have repeated the same analysis, excluding the customers that left a tip greater than 30%, the results are clearer to interpret. In this case the scatter plot and the histogram, which corresponds to the distribution around 0 of the random correlation coefficient, are the following:



The derived p-value is less than 0.01 and so the evidence against the null hypothesis is strong. We can safely reject the hypothesis of no linear relationship between the two variables and support that the generous customers are outliers in our dataset.